NORTHWESTERN UNIVERSITY

Real-time Operation of Shared-use Autonomous Vehicle Mobility Services: Modeling, Optimization, Simulation, and Analysis

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Civil and Environmental Engineering

By

Michael Francis Hyland

EVANSTON, ILLINOIS

December 2018

ABSTRACT

Real-time Operation of Shared-use Autonomous Vehicle Mobility Services: Modeling, Optimization, Simulation, and Analysis

Michael Francis Hyland

Two recent developments in the transportation industry – shared-use mobility services and fully-autonomous vehicles (AVs) – have the potential to fundamentally transform urban mobility. Shared-use mobility services (e.g. Uber, Lyft, Via, Chariot, ZipCar, and Car2go) are already beginning to bridge the gap between personal vehicles and fixed-route transit service in terms of cost, convenience, comfort, and efficiency for many trip purposes. The integration of AVs within shared-use mobility services should accelerate their growth via significantly decreasing their operational costs.

The two major advantages of AVs over human-driven vehicles within mobility service fleets are (1) the elimination of driver-related labor costs that currently make up a considerable portion of the total operational costs of ridesourcing and taxi services and (2) their ability to improve operational efficiency via allowing a central fleet operator to completely control all vehicle plans (i.e. routes, schedules, repositioning, and user assignments) in real-time. The first advantage is rather evident, and its potential impact on mobility services and future transportation systems is quite large. The second advantage (i.e. complete control of all AV plans) allows mobility service providers to optimize the operation of the entire mobility fleet with complete information on all

vehicles and traveler requests, rather than relying on individual driver decisions with incomplete information.

Given the potential paradigm shift in urban mobility due to shared-use AV mobility services (SAMSs), research in this area is particularly important. This thesis focuses on research problems associated with operating SAMSs, particularly, modeling, controlling, simulating, and analyzing the real-time operation of SAMSs. From a SAMS provider perspective, operating SAMS fleets efficiently can improve service quality, reduce operational costs, increase profitability, and increase market share. From an individual traveler perspective, assuming reasonable competition between mobility service providers, more efficient SAMS operations should lower prices and improve service quality for travelers. From a societal perspective, efficiently operating SAMS fleets can decrease (unproductive) vehicle miles thereby potentially decreasing congestion, fuel consumption, and harmful emissions.

Motivated by the importance of SAMS operational efficiency in terms of capturing benefits for mobility service providers, individual travelers, and society as a whole, as well as the inclusion of AVs in shared-use mobility services of the future, the overarching goal of this thesis is to support the operation of specific SAMS offerings via defining, modeling, and presenting solution approaches for SAMS operational problems. The specific objectives of this thesis include (1) identifying relevant SAMS operational problems, (2) analyzing the efficiency of existing taxi fleets, (3) modeling and developing solution approaches for several timely SAMS operational problems, and (4) analyzing the relative operational efficiency of specific SAMSs using agent-based stochastic simulation methods.

To meet the first objective, the thesis presents a taxonomy of vehicle routing problem variants relevant to SAMS operational problems. To address the second objective, this thesis presents two metrics to characterize the operational efficiency of a taxi fleet based on taxi trip data. The core of the thesis lies in meeting the third objective. One SAMS operational problem deals with dynamically assigning AVs to open user requests for an *on-demand SAMS without shared rides*. Another SAMS operational problem deals with simultaneously assigning AVs to open user requests and repositioning AVs throughout a service region to serve future demands, for an *on-demand autonomous carsharing service*. The last SAMS operational problem deals with assigning idle and en-route drop-off AVs to open user requests for an *on-demand shared-ride SAMS*.

The SAMS operational problems presented in this thesis represent original instances of stochastic dynamic vehicle fleet operational problems. While they share many features with problems in the existing dynamic freight routing literature, taxi-dispatching literature, and ambulance-dispatching literature, the combination of the SAMS operational problems' size, degree of dynamism, degree of urgency, spatial distribution of user requests, and short user pickup and drop-off times make the problem instances unique relative to the existing literature.

Finally, to meet the fourth objective, this thesis employs agent-based stochastic simulation methods to analyze the operational efficiency of specific SAMSs under various conditions. For example, the thesis presents a methodological framework to evaluate and quantify the impact of spatio-temporal demand forecast aggregation on the performance of an *on-demand SAMS without shared rides*. Additionally, the thesis evaluates the operational efficiency benefits of an *on-demand shared-ride SAMS* compared to an *on-demand SAMS without shared rides* from the perspective of the SAMS fleet operator.

The results of the analysis comparing an *on-demand shared-ride SAMS* to an *on-demand SAMS without shared rides* are particularly valuable from a transportation planning and policy-making perspective. The analysis indicates significant SAMS fleet operator benefits associated with offering a *shared-ride* SAMS in addition to the individual mobility benefits (significantly lower travel costs at the expense of slightly longer in-vehicle travel times) and societal benefits (more shared-ride trips implies higher vehicle occupancy which subsequently implies lower vehicle miles, traffic congestion, fuel consumption, and vehicle emissions). From a SAMS fleet operator perspective, providing a shared-ride service requires a significantly smaller fleet size than a service without shared rides, even when the maximum number of traveler groups in an AV is two and the maximum user detour distance/time is only allowed to be 5 percent more than the user's shortest route distance/time.

In meeting the above four research objectives, this thesis makes several valuable scientific contributions and provides significant value to several entities in the transportation industry. The scientific contributions range from modeling and developing solution strategies for new stochastic dynamic vehicle routing problem instances to analyzing the operational efficiency of existing taxi fleets and future on-demand SAMSs. The solution approaches can inform existing mobility service operators and future SAMS operators. The modeling framework for SAMSs can help transportation modelers incorporate SAMSs into transportation network models. Finally, the operational efficiency analyses can inform transportation planners and policy makers as they consider plans and regulations, respectively, for AVs and SAMSs.

ACKNOWLEDGEMENTS

I would like to begin by thanking my thesis committee, Prof. Hani S. Mahmassani, Prof. Karen Smilowitz, Prof. Yu (Marco) Nie, and Prof. Mike Hewitt, for their guidance and feedback throughout my doctoral candidacy. I am honored to have such a distinguished group of transportation science, operations research, and transportation systems engineering researchers on my committee.

I am especially grateful to my PhD advisor, Prof. Mahmassani, for his invaluable instruction, guidance, supervision, and mentorship. It has been an honor and a privilege to work with him for the past five years. This thesis and my research benefited greatly from his (i) experience and expertise addressing dynamic vehicle routing problems, (ii) insight into the potential economic and operational advantages of autonomous vehicle fleets, and (iii) holistic vision and knowledge of transportation systems.

I would also like to thank the other members of the transportation faculty at Northwestern University, including, Prof. Joseph Schofer, Prof. Pablo Durango-Cohen, and Prof. Amanda Stathopoulos, in addition to Profs. Mahmassani, Nie and Smilowitz. Their excellent instruction motivated many of the topics presented and methods employed in this thesis. Similarly, I owe a debt of gratitude to my undergraduate research and academic advisors, Prof. Mark Turnquist, Prof. Linda Nozick, Prof. Oliver Gao, and Dr. Francis Vanek, who introduced me to academic research and the field of transportation and encouraged me to pursue a PhD.

I want to thank the Northwestern University Transportation Center (NUTC) staff who provided valuable support for my research, including Joan Pinnell, Hillary Bean, Cynthia Ross, and Bret Johnson. I also want to thank my friends and colleagues at the NUTC, including Dr. Ali

Zockaie, Dr. Alireza Talebour, Dr. Ying Chen, Dr. Hooram Halat, and Dr. Zihan Hong for teaching me how to be a graduate student and researcher, as well as, Xiang (Alex) Xu, Archak Mittal, Amr Elfar, Marija Ostojic, Haleh Ale Ahmad, and Lama Al Hajj Hassan for their camaraderie during graduate school.

I owe a special debt of gratitude to my close friends and colleagues Dr. Andreas Frei, Dr. Omer Verbas, and Dr. Charlotte Frei who provided invaluable guidance and mentorship in addition to friendship throughout my PhD. I consider myself very lucky to have had such a helpful, caring, and intelligent group of post-doctoral and more senior graduate student colleagues who went out of their way to work with and mentor me my first two years.

Similarly, I am extremely grateful for the friendship, encouragement, and support of my close friend and colleague Dr. Lama Bou Mjahed. She is always the most gregarious and warmest person in the room, whether it be a grain elevator in North Dakota, a transportation research conference, a social event we organized, or Monday morning at the office. I also want to thank Helen Pinto for being a friend, co-author, and lunch companion these past couple years. I have found that like eating lunch, doing research is *usually* more enjoyable with close friends; hence, I have Helen, Lama, Charlotte, Omer, and Andi to thank for making both the lunch hour and the long working hours much more enjoyable.

I must acknowledge Dr. Ying Chen and Florian Dandl who were either primary authors or major contributors to journal articles associated with two chapters in this thesis.

This thesis research was funded by the U.S. Department of Transportation via the Dwight David Eisenhower Transportation Fellowship; the Northwestern University Transportation Center via its dissertation year fellowship among other sources; and the McCormick School of

Engineering via its terminal year fellowship. I want to thank all these entities for supporting this important research.

Finally, I would like to thank my family for their love and support throughout my studies. I want to thank my brother and best man James Hyland, my sister Colleen White, my sister-in-law Johanna Hyland, my brother-in-law Michael White, my brother-in-law Steven Richard, and my parents-in-law Steven and Kim Richard. I am exceptionally grateful to my parents James and Rita Hyland for their sacrifices, hard work, guidance, valuing of education, and unending love and support that, above everyone and everything else, were/are the key to any success I have had or will have in my life. I cannot thank them enough.

I am also endlessly thankful to my wife, Hillary Hyland, for her love, encouragement, friendship, support, copy editing, and so much more throughout five years of graduate school. I can always count on her to lift me up when I lack confidence and keep me in check when my ego is getting the best of me. I benefitted greatly from the sacrifices she made for me throughout graduate school. I cannot thank her enough.

DEDICATION

To Mom, Dad, and Hillary

CONTENTS

ABSTRACT 3

ACKNOWLEDGEMENTS 7

DEDICATION 10

CONTENTS 11

LIST OF FIGURES 16

LIST OF TABLES 18

Chapter 1 Introduction 19

1.1 Motivation 19

1.2 Goals and Objectives 23

1.3 SAMSs and SAMS Operational Problems 24

1.4 Contributions 28

1.5 Organization 30

Chapter 2 Taxonomy 32

2.1 Overview 32

2.2 Classifying SAMS Operational Problems 35

Pickup and/or Drop-off 35

Evolution and Quality of Information 36

Availability of Information 39

Processing of Information 39

Time-Window Constraints 40

Fleet Size 41

2.3 Taxonomic Categories to Classify SAMS Operational Problems 42

Vehicle Homogeneity 42

Vehicle Capacity Constraints 43

Maximum Vehicle Route Times and Distances 43

Costs 44

Objective ... 45
2.4 Taxonomic Categories Highly Relevant to SAMS Operational Problems ... 46
Fleet Size Elasticity ... 47
Reservation Structure ... 47
Pricing ... 48
Accept/Reject Decision ... 49
Reservation Timeframe (Degree of Dynamism) ... 50
Shared Rides ... 52
Repositioning of Vehicles ... 53
Underlying Network ... 54
Network Congestion/Travel Times ... 55
Demand Reassignments/Vehicle En-route Diversions ... 55
Decision Epochs ... 56
2.5 Conclusion ... 56
Chapter 3 Literature Review ... 57
3.1 Overview ... 57
3.2 Degree of Dynamism ... 58
3.3 SDVRPs Applications ... 60
Taxi dispatching ... 60
Ambulance Dispatching and Fleet Management ... 61
Freight Dynamic Truckload Pickup and Delivery ... 63
Mobility-on-demand (MOD) and autonomous-MOD (AMOD) Services ... 65
Shared Rides ... 68
3.4 Conclusion ... 70
Chapter 4 Operational Efficiency of Chicago Taxi Fleet ... 71
4.1 Overview ... 71
4.2 Chicago Taxi Trip Data ... 72
4.3 Taxi Fleet Usage Analysis ... 74

Histogram of Taxi Fleet by Average Daily Trips 74
Clustering Taxis by Daily Trip Count 75
4.4 Taxi Efficiency Analysis 82
Efficiency Metrics 82
Spatial Efficiency Metric 85
Google Maps Distance Matrix API 86
Efficiency Results 87
4.5 Conclusion 92
Chapter 5 Modeling Framework 93
5.1 Overview 93
5.2 Model Components 93
Decision Epochs 94
State Variables 94
Decision Variables 94
Exogenous Information 95
Transition Function 95
Objective Function 96
5.3 Conclusion 96
Chapter 6 On-demand SAMS without Shared Rides 97
6.1 Overview 97
6.2 Motivation and SAMS Definition 98
6.3 Problem Statement and Model 100
Problem Statement 100
Model 101
6.4 Solution Approach 106
AV-User Assignment Problem 106
AV-User Assignment Strategies 109
6.5 Experiments and Computational Results 117

Artificial Demand ... 119

Chicago Taxi Demand ... 126

6.6 Conclusion ... 128

Chapter 7 Assessing the Operational Benefits of Shared-Ride SAMSs ... 130

7.1 Overview ... 130

7.2 Introduction ... 131

7.3 Research Methodology ... 133

User Requests ... 134

On-demand SAMS without Shared Rides Fleet Operations ... 134

On-demand Shared-ride SAMS Fleet Operations ... 134

7.4 Experimental Design ... 138

7.5 Results ... 139

7.6 Conclusion ... 143

Chapter 8 On-demand Autonomous Carsharing Service ... 145

8.1 Overview ... 145

8.2 Motivation ... 146

8.3 ODACS Problem ... 150

Problem Statement ... 151

Model ... 152

8.4 Joint Assignment-Repositioning Operational Policy ... 159

8.5 Experimental Design ... 161

User Request Data ... 161

Parameter Settings and Scenarios ... 162

Other Operational Strategies ... 164

8.6 Computational Results ... 164

8.7 Conclusion ... 168

Chapter 9 Impact of Spatio-Temporal Demand Forecast Aggregation on the Operational Performance of Shared Autonomous Mobility Fleets ... 170

9.1 Overview ... 170

9.2 Motivation ... 171
9.3 Background ... 172
9.4 Research Problem and Hypothesis ... 174
9.5 Research Methodology ... 175
Agent-based Simulation Framework ... 176
User Requests ... 178
Demand Forecasts ... 178
SAMS Fleet Operational Strategy ... 180
Fleet Strategy without Demand Forecasts ... 185
9.6 Experimental Design ... 185
NYC Taxi Data ... 185
Parameter Settings ... 190
Scenarios ... 192
9.7 Results ... 193
Demand Forecasting Results ... 193
SAMS Fleet Performance Results ... 195
9.8 Conclusion ... 200
Summary and Implications ... 200
Limitations and Future Work ... 201
Chapter 10 Concluding Remarks ... 203
10.1 Summary and Contributions ... 203
10.2 Applications ... 205
10.3 Future Research Areas ... 205
REFERENCES ... 208

LIST OF FIGURES

Figure 4-1: Distribution of the average number of daily trips (per in-service day) across taxis .. 75

Figure 4-2: Taxi clusters results 79

Figure 4-3: Histograms for the percentage of days in service for each taxi cluster...................... 81

Figure 4-4a: Algorithm to obtain the productive time of Taxi-trip $\boldsymbol{j}, \boldsymbol{n}$. 83

Figure 4-5a: Algorithm to obtain the loaded distance of taxi-trip $\boldsymbol{j}, \boldsymbol{n}$... 85

Figure 4-6: Histograms of the average temporal efficiency of taxis in the most (top) and normal (bottom) trips clusters by year 89

Figure 4-7: Histograms of the spatial efficiency of taxi #44 (top) and taxi #981’s (bottom) trips from 2013 to 2016 91

Figure 6-1: AV-user assignment based on Strategy 1 (total fleet miles = 12.9)......................... 111

Figure 6-2: AV-user assignment based on Strategy 2 (total fleet miles = 9.7)........................... 112

Figure 6-3: AV-user assignment based on Strategy 3 (total fleet miles = 9.0)........................... 112

Figure 6-4: AV-user assignment based on Strategy 4 (total fleet miles = 6.8)........................... 114

Figure 6-5: AV-user assignment based on Strategy 5 (total fleet miles = 7.4)........................... 115

Figure 6-6: AV-user assignment based on Strategy 6 (total fleet miles = 5.4)........................... 117

Figure 7-1: Average user in-vehicle travel time as a function of demand (x-axis) and maximum user detour time/distance (line color) 140

Figure 7-2: Average user wait time as a function of demand (x-axis) and maximum user detour time/distance (line color) 141

Figure 7-3: Total fleet miles as a function of demand (x-axis) and maximum user detour time/distance (line color) 142

Figure 7-4: Percentage of users who shared a ride as a function of demand (x-axis) and maximum user detour time/distance (line color) 143

Figure 8-1: Comparison of operational strategies (line color) in terms of average user wait time (top), total empty fleet kilometers (middle) and system cost (bottom) across fleet sizes (x-axis) 167

Figure 9-1: Schematic of process to convert data into better decisions..................................... 172

Figure 9-2: Simulation framework.. 177

Figure 9-3: Subregion layout for different spatial aggregation levels (official NYC taxi subregions for Manhattan are drawn in the background) 187

Figure 9-4: Taxi trip density during the morning (a-c) and evening (d-f) on Wednesday, 2016-04-06 for trip origins (a and d), trip destinations (b and e), and net trips (c and f) wherein red (green) areas represent more (fewer) trip origins, trip destinations, and net trips, respectively 189

Figure 9-5: Average demand forecast errors for different spatial (x-axis) and temporal (line type) aggregation levels according to RSMRE (top) and sMAPE (bottom) error measures........ 194

Figure 9-6 Average user wait time as a function of subregion edge length (x-axis) and forecast type (line color)... 197

Figure 9-7 Empty fleet miles as a function of subregion edge length (x-axis), type of empty miles (line type), and forecast type (line color)... 198

LIST OF TABLES

Table 2-1: Taxonomy of Vehicle Routing Problems 34
Table 4-1: Chicago Taxi Trip Dataset General Statistics 73
Table 4-2: Individual Taxi Trip Statistics 74
Table 4-3: Percentage of Taxis in Each Cluster 77
Table 6-1: Overview of AV-User Assignment Strategies 117
Table 6-2: Parameter Values for Artificial Demand Scenarios 120
Table 6-3: Average Travel Time (min) Results for Uniform Demand Scenarios 122
Table 6-4: Ratio of Empty Fleet Miles to Total Fleet Miles Results for Uniform Demand Scenarios 124
Table 6-5: Computational Results for the Clustered Artifical Demand Scenarios 126
Table 6-6: Experimental Results for Taxi Demand 128
Table 7-1: Parameter Values for Experiments 139
Table 8-1: Parameter Settings that do not Vary across Scenarios 163
Table 9-1: Parameter Settings that do not Vary across Scenarios 192
Table 9-2: Complete SAMS Fleet Performance Results as a Function of Forecast Type, Spatial Aggregation, and Temporal Aggregation 199

Chapter 1 Introduction

1.1 Motivation

This thesis is motivated by two recent developments in the transportation industry, namely, the emergence and growth of shared-use mobility services as well as the advent of fully-autonomous vehicles (AVs). Together AVs and shared-use mobility services – such as those provided by ridesourcing/ride-hailing companies Uber and Lyft, (and to a lesser extent) carsharing companies Zipcar and Car2go, and micro-transit companies Chariot and Via – have the potential to significantly alter existing passenger transportation systems and improve urban mobility.

The individually-owned and -operated (i.e. personal) vehicle has been the dominant mode of passenger transportation in most U.S. cities for over sixty years, accounting for over 75% of commute travel (Mckenzie and Rapino, 2011) and resulting in significant traffic congestion in metropolitan areas across the United States. Fixed-route transit also serves a sizeable portion of trips in many large metropolitan areas, including New York-Northern New Jersey (30.5% of commute travel), San Francisco-Oakland (14.6%), and Washington D.C. (14.1%). However, for many travelers and trip purposes in many regions, there is a large gap between these two modes in terms of cost (fixed and marginal), convenience, comfort, and efficiency. Personal vehicles are much more expensive to own and operate than using transit, but they usually provide superior convenience, comfort, and efficiency for most travelers on most trips[1].

Shared-use mobility services, which have gained significant market share over the past five-to-ten years (Clewlow and Mishra, 2017), are beginning to bridge the gap between personal vehicles and fixed-route public transit. However, for most trips and trip purposes, the personal vehicle or fixed route transit is still cheaper for travelers and/or more convenient and efficient than shared-use mobility services. Fortunately, AVs have the potential to change the underlying economics and significantly improve the operational efficiency of shared-use mobility services via eliminating the costs and performance limitations of human drivers, respectively. These reductions in operational costs and improvements in operational efficiency should allow shared-use AV mobility services (SAMS[2]s) to capture a sizeable portion of trips in many metropolitan regions.

A shift away from personal vehicle usage towards SAMSs has significant implications on transportation systems. If SAMS travelers continue to follow the same travel patterns and do not share rides with other travelers, total vehicle miles travelled (VMT) on roadways will likely increase due to the increase in empty VMT needed for AVs to travel between the drop-off location of one traveler and the pickup location of next traveler. Moreover, if travel becomes cheaper (in terms of both productive time lost during travel and monetary cost), some people may decide to move farther away from activity (e.g. work, shopping, and entertainment) locations and travel longer distances thereby further increasing total VMT. Alternatively, if SAMS users are willing to

[1] There are certainly trips that are cheaper, more convenient, and/or more efficient on transit than via a personal vehicle; however, in this U.S. these trips are relatively rare because of personal housing and travel preferences as well as the land-use and transportation infrastructure investment decisions at all levels of government for decades.

[2] In this dissertation, a SAMS is a general term describing a passenger transportation/mobility service provided by a private company using a fleet of privately-owned and -operated fully-autonomous vehicles.

share rides, and SAMS providers offer shared-ride services, total VMT may decrease due to higher vehicle occupancies. Moreover, without the need to own and store (i.e. park) a personal vehicle at home, people may decide to live in denser urban areas where they can more easily access employment opportunities, entertainment, shopping, and other activities without needing to use a vehicle.

AVs alone should improve road safety via eliminating human drivers; they should also increase the capacity and traffic stability of highways (Talebpour and Mahmassani, 2016), thereby, potentially reducing traffic congestion, fuel consumption, and harmful emissions. The inclusion of AVs within shared-use mobility service fleets offers additional potential benefits. For example, SAMSs can potentially:

- Reduce overall transportation costs, especially for families who currently own and operate multiple personal vehicles. First, SAMSs can nearly eliminate parking costs. Second, initial research suggests one shared AV can replace several (up to 10) personal vehicles. This means that the high purchasing, maintenance, and insurance costs of a vehicle can be split among several SAMS users. Third, if users are willing to share rides, they can split the operational costs, mainly fuel, of traveling in a vehicle.

- Enhance mobility and accessibility, especially for involuntary car-less travelers. SAMS can provide affordable and efficient transportation between origin-destination pairs that are poorly served via existing transit systems.

- Decrease overall VMT, traffic congestion, fuel consumption, and vehicle emissions. This, once again, depends on SAMS providers offering shared-ride services and SAMS users willing to share rides. Shared-ride service increases vehicle occupancy thereby requiring fewer vehicles to transport the same number of people.

To help capture the benefits of SAMSs, this thesis addresses research challenges surrounding the operation of SAMSs. Operational efficiency is a key component in the success of any transportation system including SAMSs. Inefficient SAMS operations can engender poor service quality, high operational costs, and/or increase fleet size requirements thereby making SAMSs less attractive in terms of affordability and service quality. This thesis addresses on-demand SAMSs with and without shared-rides.

AVs should provide significant operational advantages to SAMSs relative to human-driven vehicles. The most relevant operational-level advantage of AVs, over human-operated vehicles, is their ability to safely and near-instantaneously receive and execute changes in vehicle plans (e.g. routes, schedules, and user assignments) coming from the fleet controller who has near-perfect information about all the AVs and traveler requests in the system and can make decisions that optimize the entire fleet's performance. The computational experiments in this thesis illustrate the efficiency advantages associated with operational strategies that allow fleet controllers to frequently change the plans of individual vehicles in real-time as new user requests are made.

From a fleet management perspective, the biggest advantage of AVs is their guaranteed compliance with these real-time plan changes, and more generally, the fleet manager's operational policies. Although it is possible to force taxi and ridesourcing drivers to follow the fleet manager's operational policies and the fleet controller's real-time instructions, the fact that taxi and ridesourcing services give drivers considerable autonomy suggests that driver compliance may be too difficult to mandate and/or ineffective in practice. With complete operational control, the fleet controller can optimize the operations of the entire fleet rather than having individual agents (i.e. human drivers) try to optimize their own objectives. In practice, this may require some AVs to wait in areas with low demand, while other AVs shuttle users back-and-forth between high-demand areas. While taxi drivers would almost certainly not accept this operational strategy, robot vehicles will not have any qualms with it.

1.2 Goals and Objectives

Motivated by the potential societal and individual mobility benefits of SAMSs and the importance of operational efficiency in terms of capturing these benefits, the overarching goal of this thesis is to enable the efficient operation of SAMSs through developing models and solution approaches for SAMS operational problems and analyzing the efficiency of specific SAMSs.

Specifically, the objectives of this thesis include (1) identifying relevant SAMS operational problems, (2) developing methods and an evaluation methodology to analyze the efficiency of existing taxi fleets to subsequently motivate the importance of operational efficiency in SAMSs, (3) modeling and developing solution approaches for several timely SAMS operational problems, and (4) using agent-based stochastic simulation methods to analyze the relative operational efficiency of specific SAMSs.

To meet the first objective, this thesis presents a taxonomy of vehicle routing problem variants relevant to the operation of SAMSs. This thesis also presents a wide-ranging, yet in-depth, review of the existing literature that relates to SAMS operational problems. Broadly, the taxonomy aims to support research related to the operation and management of SAMS fleets. Specifically, it aims to provide researchers and analysts examining SAMS operational problems a valuable reference to identify relevant problem classes and problem variants.

To meet the second objective, this thesis presents two metrics to characterize the operational efficiency of a taxi fleet based on taxi trip data. The two metrics are a spatial efficiency metric and a temporal efficiency metric which approximate the percentage of unproductive fleet miles and unproductive time, respectively, for the taxi fleet. This thesis quantifies the operational efficiency of the Chicago taxi fleet using these two metrics.

The core of the thesis lies in meeting the third and fourth objectives. The next subsection presents three SAMSs, defines three SAMS operational problems, and identifies two important research questions relating to the operational efficiency of specific SAMSs.

1.3 SAMSs and SAMS Operational Problems

This thesis defines and models operational problems associated with three SAMSs and presents operational strategies (i.e. policies) to solve the stochastic dynamic fleet operational problems associated with these SAMSs. The three SAMSs are the *on-demand SAMS without shared rides*, the *on-demand shared-ride SAMS*, and the *on-demand autonomous carsharing service (ODACS)*. For all three SAMSs:

- The AVs are owned and operated by a mobility service provider
- Users request rides via their smartphone
- Users want to be picked up immediately and expect to be picked up within a few minutes
- User requests consist of a pickup location and a drop-off location

- The AVs in the fleet are functionally homogenous

- The AV fleet size is fixed.

The *on-demand SAMS without shared rides* is essentially a ridesourcing/ride-hailing (e.g. UberX or traditional Lyft) service, except (i) the vehicles are driverless, (ii) fleet size is fixed, and (iii) the SAMS provider owns, and has complete control over the AVs. The AV fleet transports users directly between their requested origin locations and destination locations; i.e. there are no detours to pick up or drop off other users.

As the name suggests, an *on-demand shared-ride SAMS* is essentially the same as the *on-demand SAMS without shared rides* except shared-rides are allowed. Ceteris paribus, a share-ride SAMS provider will offer travelers lower prices at the expense of longer in-vehicle travel times and having to share a vehicle with other travelers.

The *on-demand autonomous carsharing service (ODACS)* is essentially a carsharing service (e.g. ZipCar, Car2go), except (i) the vehicles are driverless, (ii) they can reposition themselves throughout the service region, and (iii) they can pick users up at their points of origin. In addition to an origin location and destination location, user requests include an estimated usage time, which is a conservative estimate of the time the user will need an AV before releasing it to the fleet controller at the drop-off location. The fleet controller cedes control of AV plans (i.e. where to go next) to users once the AVs arrive at the users' pickup locations.

This thesis defines the operational problems for all three SAMS, models their underlying stochastic dynamic operational problems, and presents solution approaches. The first SAMS operational problem this thesis addresses relates to the *on-demand SAMS without shared rides*. To focus on the assignment of AVs to user requests, this specific problem assumes the SAMS provider does not have deterministic or stochastic information about future user requests, other than an estimate of the demand rate for the whole region. Hence, the fleet controller must find an operational strategy/policy to assign AVs to users as they request service dynamically. This is a highly-dynamic and stochastic problem as user requests arrive in real-time and users want to be picked up immediately. To address this problem, this thesis presents and tests several AV-user dynamic assignment strategies.

The second SAMS operational problem this thesis addresses relates to the *on-demand shared-ride SAMS*. Rather than focusing on the underlying operational problem associated with the *on-demand shared-ride SAMS,* the thesis aims to illustrate the operational efficiency benefits of allowing shared rides via comparing the *on-demand shared-ride SAMS* with the *on-demand SAMS without shared rides*. Given a fixed fleet size, the computational analysis compares the two SAMSs in terms of their ability to serve different levels of traveler demand. This chapter also presents a sensitivity analysis illustrating the impact of maximum user in-vehicle detour distance/time on the operational efficiency of the AV fleet.

The third SAMS operational problem this thesis addresses relates to the *on-demand autonomous carsharing service (ODACS).* In this problem, the ODACS fleet controller does have stochastic (i.e. historical) information about the spatio-temporal distribution of user requests. Typical models and solution approaches to this problem divide the operational problem into an AV-user assignment/dispatching subproblem and an empty AV repositioning subproblem and use heuristic methods for one or both subproblems. This chapter models and presents a solution approach to the problem that simultaneously assigns AVs to open user requests and repositions empty AVs to subregions of the service area with an expected imbalance between supply and demand. In addition to stochastic user requests, the specific problem in this thesis includes a second element of stochasticity, namely, the actual AV usage time of carshare users. While users request access to an AV in increments of fifteen minutes, and they are incentivized to make conservative requests, their actual usage and therefore AV drop-off times are uncertain. Given some of the operational strategies in this study allow the fleet controller to assign currently in-use AVs to user requests, the operator must consider and estimate the remaining user usage time based on the user's requested usage time and the known distribution of the difference between requested and actual usage times. For this stochastic dynamic operational problem, the fleet controller's objective is to minimize user wait times and minimize empty AV fleet miles. To address this operational problem, this thesis presents a joint assignment-repositioning operational strategy. The joint assignment-repositioning formulation captures the immediate rewards, immediate costs, and costs-to-go given the system state at each decision epoch.

This thesis also addresses an important research problem relating to the operation of an *on-demand SAMS without shared rides,* wherein the fleet controller has stochastic information about future user requests. The research problem is to quantify the impact of spatio-temporal demand forecast aggregation on the performance of an *on-demand SAMS without shared rides* fleet. In general, when short-term forecasts of user requests are intended for a finer space-time discretization, they tend to become less reliable due to a smaller average number of requests per space-time bin and the law of large numbers. However, holding forecast quality constant, more disaggregate forecasts provide more valuable information to fleet controllers. For example, knowing three users will request rides in a 100 m^2 area between 9:00am and 9:05am is more valuable than knowing three users will request rides in a 1000 m^2 area between 9:00am and

9:30am. To address this tradeoff, this thesis presents a flexible methodological framework that includes an agent-based simulation model, a short-term demand forecasting method, and an SAMS fleet operational strategy that incorporates short-term demand forecasts. The analysis involves simulating dozens of scenarios to evaluate the impact of spatio-temporal demand forecast aggregation on SAMS fleet operational performance.

1.4 Contributions

This thesis makes several conceptual and methodological contributions to the scientific literature, as well as several valuable computational findings. First, this thesis presents a taxonomy of vehicle routing and vehicle fleet operational problems from the literature in order to inform future research on SAMS operational problems. The existing literature on vehicle routing problems is wide-ranging and extensive; hence, a characterization of the important problem elements, relevant to SAMS operational problems, should provide researchers a valuable resource when tackling new problem classes and problem instances.

Second, this thesis clearly defines three SAMSs and their operational problems (see Section 1.3 for a description of the three SAMSs and their operational problems). Given the high likelihood of these three SAMS offerings in the future, it is important to identify, define, model, and present solution approaches for their operational problems.

Third, this thesis presents a comprehensive analysis/comparison of operational policies for the dynamic assignment of AVs to user requests, in an *on-demand SAMS without shared rides*, in the case where the fleet controller has no spatial information about future user requests. The thesis compares six different AV-user assignment strategies including a dynamic AV-user assignment strategy that considers all AVs and all unassigned and assigned (but not yet picked up) user requests at each decision point. This approach allows assigned AVs to be diverted (and users to be reassigned) to different user requests (to different AVs). Moreover, the dynamic AV-user assignment strategy also allows en-route drop-off AVs to be assigned to open user requests. This allows current in-use/en-route drop-off AVs to be assigned to open user requests. This approach unambiguously outperforms the other operational strategies in terms of empty/unproductive fleet miles.

Fourth, this thesis compares the operational efficiency of an *on-demand shared-ride SAMS* with an *on-demand SAMS without shared rides*. This comparison is particularly relevant for transportation planning and policy purposes. Results illustrate the superiority of a *shared-ride SAMS* in terms of handling high demand levels with a limited vehicle fleet. As demand increases, unlike the *SAMS without shared rides*, the *shared-ride SAMS* effectively increases it service rate (customers served per hour) due to economies of density. Shared-ride fleet users experience slightly longer in-vehicle travel times but typically incur shorter wait times and will presumably receive lower fares due to a decrease in fleet miles (operational costs) and fleet size (capital costs) compared with non-shared-ride fleet users.

Fifth, this thesis presents a joint assignment-repositioning strategy to address the ODACS operational problem. The joint assignment-repositioning formulation allows the fleet controller to implicitly make trade-offs between assigning AVs to open requests now, reducing subregion imbalances now, and waiting until later (when other AVs become available) to assign AVs to open user requests or balance subregions. The results indicate this strategy to be very effective in terms of reducing average user wait times relative to myopic operational strategies that do not incorporate repositioning.

Sixth, this thesis includes a research problem relating to the integration of short-term demand forecasts and SAMS operational strategies, as well as a flexible methodological framework to address the research problem. This thesis aims to evaluate and quantify the impact of spatio-temporal demand forecast aggregation on the performance of an on-demand SAMS fleet. The flexible methodological framework includes an agent-based simulation model with a short-term demand forecast module and an SAMS fleet operational strategy that incorporates short-term demand forecasts module. The analysis suggests that more disaggregate forecasts (i.e. smaller subregions) produce better results. However, compared to scenarios with perfect spatio-temporal demand forecasts, improvements begin to plateau when using an offline historical average model to forecast user requests.

1.5 Organization

Chapter 1 motivates the thesis topic, presents the overarching goal and specific objectives, conceptualizes the key problems, and describes the scientific contributions of the thesis. Chapter 2 includes a taxonomy of vehicle routing problems relevant to SAMS operational problems. Chapter 3 provides a review of the academic literature relevant to on-demand SAMS operational problems. Chapter 4 presents spatial and temporal efficiency metrics to evaluate the operational efficiency of taxi fleets and mobility services. Using these metrics and taxi trip data, this chapter analyzes the operational efficiency of the Chicago taxi fleet. Chapter 5 presents a framework to model on-demand SAMS operational problems that can be classified as stochastic dynamic vehicle routing/fleet operational problems. Chapter 6 focuses on the problem of assigning (or matching) AVs to open user requests in real-time for an *on-demand SAMS without shared rides*. Chapter 7 compares the operational efficiency of an *on-demand shared-ride SAMS* and an *on-demand SAMS without shared rides*. Chapter 8 focuses on operating an *on-demand autonomous carsharing services* via assigning AVs to user requests and simultaneously repositioning AVs to serve future demands. Chapter 9 examines the impact of spatio-temporal demand forecast aggregation on the operational efficiency of an *on-demand SAMS without shared rides*. Chapter 10 concludes the thesis and discusses areas of future research.

Chapter 2 Taxonomy[3]

2.1 Overview

The SAMS operational problems presented in this thesis fall under the umbrella of vehicle routing problems (VRPs). The existing literature on VRPs is extensive and diverse; hence, this taxonomy chapter aims to assist analysts and researchers in identifying the type of problem they are confronting. Once the analyst or researcher is able to identify the characteristics of the problem in question, she can either make use of existing problem formulations and solution algorithms from the literature, if they exist, or begin to define a new problem class and develop original solution algorithms (Bodin and Golden, 1981, p. 98).

This chapter classifies the general SAMS operational problem using existing taxonomic categories in the literature. Then, to classify specific SAMS operational problems, this chapter adds additional, or more nuanced, dimensions to existing taxonomic categories. Finally, this chapter presents new taxonomic categories to further classify relevant SAMS operational problems.

Table 2-1 displays a classification of VRPs. SAMS operational problems can be broadly classified as dynamic, multi-vehicle, pickup and delivery problems with hard (explicit) or soft (implicit) time-window constraints. Highly-relevant categories presented include variable/fixed

[3] This chapter extends the research published in Hyland and Mahmassani (2017).

fleet size, reservation structure, accept/reject decision-maker, reservation timeframe, vehicle repositioning, underlying network structure, sharing of rides, and network congestion. The two goals of the taxonomy are to provide researchers a valuable reference as they begin to model SAMS operational problems and to identify new SAMS operational problem classes and problem instances to spur interest from researchers.

Table 2-1: Taxonomy of Vehicle Routing Problems

Existing Taxonomic Categories		**Categories Highly Relevant to SAMS Operational Problems**
Classify General SAMS Problem	**Classify Specific SAMS Problems**	
Pickup and/or Drop-off • Pickup only • Drop-off only • Paired pickup and drop-off • Unpaired pickup and/or drop-off **Evolution of Information** • Static • Dynamic **Availability of Information** • Global • Local **Time-Window Constraints** • No time-windows • Hard time-windows • Soft time-windows • Hard and soft **Size of Vehicle Fleet** • One vehicle • Multiple vehicles	**Quality of Information** • Deterministic • Stochastic **Processing of Information** • Centralized • Decentralized **Vehicle Homogeneity** • Homogenous • Heterogeneous **Vehicle Capacity Constraints** • Imposed all the time • Imposed some of the time • Not imposed **Maximum vehicle route times (and distances)** • Imposed – all the same • Imposed – not all the same • Not imposed **Costs** • Variable or routing costs • Fixed operating or vehicle acquisition costs (capital costs) **Objective** • Maximize profit • Minimize cost • Minimize user inconvenience • Minimize vehicle miles traveled • Minimize user wait time • Minimize user in-vehicle travel time • Minimize number of vehicles • Mixed (i.e. multi-objective)	**Fleet Size Elasticity** • Elastic • Fixed Fleet Size **Reservation Structure** • Short-term rentals • Point-to-point service • Mixed **Pricing** • No pricing • Fixed pricing structure • Pricing, with no fixed structure **Accept/Reject Decision** • No decision • Fleet manager decision • Customer decision **Reservation Timeframe** • Immediate requests • Minimum pre-reservation time • Mixed **Shared Rides** • No shared rides • Shared rides **Repositioning of Vehicles** • No repositioning • Repositioning **Underlying Network** • Real road network • Test road network • Graph/Virtual Network **Network Congestion/Travel Times** • Static and Deterministic • Time-dependent and Deterministic • Time-dependent and Stochastic **Reassignments/ En-route Diversions** • Allowed • Not allowed **Decision Epochs** • One at beginning (deterministic) • Exogenous • Endogenous

2.2 Classifying SAMS Operational Problems

Research on the VRP stretches back over six decades (Dantzig and Ramser, 1959). Studies range from highly-theoretical to applications of a single problem instance. Hence, it is not surprising that other researchers have presented taxonomies of vehicle routing problems for various purposes. This section describes some of the taxonomic categories in the existing literature that are relevant to SAMS operational problems. It begins with the most important categories that draw clear distinctions across problem classes and clearly distinguish SAMS operational problems from existing VRPs in the literature.

Pickup and/or Drop-off

In their taxonomy of vehicle routing and scheduling problems, Bodin and Golden (1981) define an *operations* category with the following elements:

- Pickups only
- Drop-offs only
- <u>Mixed</u>

The pickups only and drop-offs only options refer to pure vehicle routing problems (VRPs); whereas, the mixed option is better known as the pickup and delivery problem (PDP), which is a generalization of the VRP. In the conventional PDP, each customer request has a unique pickup and drop-off location pair (i.e. a demand must travel from its given origin to its given destination). An alternative 'mixed' PDP is the unpaired PDP where a fleet of vehicles must transport goods from supply locations to demand locations, but each supply location is not paired with a demand location. This problem arises in retail redistribution.

The category name, operations, is too broad for classifying vehicle routing/fleet operational problems today. Hence, this thesis renames the category *Pickup and/or Drop-off.*

SAMS operational problems can definitively be classified as paired pickup and delivery problems. In passenger transportation in urban areas, the pickup and drop-off locations of traveler requests are unique location pairs spread across the urban area.

Evolution and Quality of Information

Pillac et al. (2013) classify VRPs based on two dimensions simultaneously, [1] evolution of information, and [2] quality of information. The evolution of information category classifies problems as *static* or *dynamic*; whereas, the quality of information category classifies problems as *deterministic* or *stochastic*. The four possible combinations of problems based on this classification pair are described below:

Static and deterministic: All the problem information (i.e. the problem inputs) including customer locations, arc travel times, node service times, etc. are known exactly and are available far enough in advance such that vehicle routes and schedules can be formed prior to the routing process. Once a solution to the problem is found, it is not adjusted during the routing process. In their seminal work, Dantzig and Ramser (1959) formulate a static and deterministic VRP. Laporte (1992) reviews problem formulations and solution algorithms for the static and deterministic VRP. Toth and Vigo (2002) edited a book on the vehicle routing problem that covers exact and heuristic solution methods to a variety of static and deterministic VRPs.

Static and stochastic: In the static and stochastic case, at least one problem input is only partially known as a random variable (Pillac et al., 2013). Possible stochastic problem inputs include arc travel times, node service times, or location of customers. The static and stochastic inputs are available far enough in advance such that vehicle routes and schedules can be formed prior to the routing process. Once the routing process begins only minor changes to the vehicle routes are allowed, and these changes do not require communication between individual vehicles/drivers and a central operator. Gendreau et al. (1996) present a review of static and stochastic VRPs. Eksioglu et al. (2009) present a taxonomy of VRP studies and categorize specific sources of stochasticity including customer service demand quantity, request times of new customers, and on-site service/waiting times.

Dynamic and stochastic: In the dynamic and stochastic VRP, necessary problem inputs reveal themselves in real-time. However, the stochastic distributions of these problem inputs can be exploited by the fleet operators when routing and repositioning vehicles. For example, if the exact location and time of future demand requests are unknown, but the spatio-temporal distribution of user requests is known, the fleet controller can take advantage of this information in the routing process. Real-time location information is necessary to solve the dynamic problem because routes and assignment decisions need to be re-calculated in real-time.

Dynamic and deterministic: There is debate in the literature about the meaning of the term *dynamic and deterministic* VRP. There is agreement that in *dynamic and deterministic* problems, the fleet controller does not have distributional information about inputs to exploit. However, in the definition of Pillac et al. (2013), in *dynamic and deterministic* problems, exact information reveals itself to the fleet controller during the routing process (Bektas et al., 2014). According to others, a *dynamic and deterministic* problem is the same as a sequential deterministic problem (Lahyani et al., 2015; Ulmer, 2017). In a sequential deterministic problem, decisions in one period impact the decisions in later periods; however, the impact of the earlier decision on the later decision, and the evolution of the system are deterministic. Hence, the problem can be modeled as a single decision problem prior to the routing process. Ulmer (2017) and Lahyani et al. (2015) view the *dynamic and deterministic* problems defined by Pillac et al. (2013), as stochastic dynamic problems without a distribution for the stochastic element. In fact, they state that only stochastic problems can be dynamic, and deterministic problems cannot be dynamic.

SAMS operational problems are inherently dynamic as most information reveals itself in real-time, such as, user pickup and delivery times and locations, arc travel times, user service times, etc. This thesis refers to all the on-demand SAMS operational problems as stochastic dynamic vehicle routing problems (SDVRPs). However, since this section uses the definition in Pillac et al. (2013), Table 2-1 says SAMS operational problems are dynamic but either deterministic or stochastic.

In the literature, the dynamic problem is also referred to as the real-time problem (Yang et al., 2004) and the online problem (Jaillet and Wagner, 2006; Yang et al., 1999). Compared with the static problem wherein a decision problem only needs to be solved once, in the dynamic problem the fleet controller needs to constantly make decisions as new information enters the system. The stochastic dynamic vehicle routing problem is still a very active research area (Marlin Ulmer et al., 2017; Ulmer, 2016)

Savelsbergh & Sol (1995) point out that in the dynamic problem, the notion of vehicle depots is not applicable because, as new customer requests arrive in real-time, and a decision problem needs to be re-solved, vehicles are located throughout the operational area, not at fixed depots.

Availability of Information

In addition to evolution and quality of information, Eksioglu et al. (2009) include two other subcategories under the information characteristics category in their VRP taxonomy including availability of information and processing of information. The availability of information taxonomic elements are:

- Global
- Local

For SAMS operational problems, wherein a fleet controller operates all the AVs in the fleet and receives information about all user requests, it is reasonable to assume the fleet controller has global information. Conversely, in existing taxi and ridesourcing services, individual drivers may only receive local information about open requests; i.e. they may only see requests within a given radius of their current location.

Processing of Information

The processing of information taxonomic elements in Eksioglu et al. (2009), are:

- Centralized
- Decentralized

Given the computational complexity of routing and scheduling vehicles dynamically, decentralized, as well as centralized computing architectures are both likely to be seen in practice. Smaller fleets may be able to employ a single centralized computing system; whereas, larger fleets may require a decentralized computing architecture. For example, each of the five boroughs in

New York City may have their own fleet controller that controls the vehicles currently in their borough. These fleet controllers still need to pass information, but they only control the AVs in their borough.

The SAMS operational problems addressed in later chapters of this thesis assume centralized information processing and decision making.

Time-Window Constraints

Bodin and Golden (1981) define a vehicle routing '*time to service a particular node or arc'* category with the following elements:

- Time specified and fixed in advance (pure vehicle scheduling problem)
- Time windows (combined vehicle routing and scheduling problem)
- Time unspecified (in this case, it is a vehicle routing problem unless there are precedence relationships as well, in which case it is a combined vehicle routing and scheduling problem)

This thesis repurposes and renames this category and in doing so removes the pure vehicle scheduling problem option and reduces the ambiguity of the time unspecified option. The updated category, *time-window constraints*, includes the following elements:

- No time-windows
- Hard time-windows
- Soft time-windows
- Hard and soft time-windows

In this classification system, the no time-windows problem allows the fleet controller to serve user requests at any time, whereas the problems with hard (explicit) and/or soft (implicit) time-windows force the fleet operator to serve user requests within specified time-windows. The difference between hard and soft time-windows can manifest itself in the mathematical formulation of the decision problem. Hard time-window constraints require inequality constraints in the mathematical program; whereas, soft time-window constraints typically influence the objective function. Lagrangian relaxation can be used to convert hard time-window constraints into soft time-window constraints. Soft time-window constraints in the objective function are often referred to as the *quality of service* term. Eksioglu et al. (2009) present a similar classification category entitled *time window structure* with the following elements, soft-time windows, strict time windows, mix of both. The pickup and delivery problem with time-windows (PDPTW) is a well-known problem with passenger and freight transportation applications (Berbeglia et al., 2010; Lu and Dessouky, 2006; Ropke and Cordeau, 2009).

SAMS operational problems are combined vehicle routing and scheduling problems. The problem can be formulated with hard and/or soft time-window constraints depending on the problem instance. For example, in 2016, ridesourcing companies Uber and Lyft implicitly operated with soft time-window constraints (customers are picked up as soon as possible); whereas, other mobility service companies may provide their customers with hard time-windows.

Fleet Size

Bodin and Golden (1981) define a *size of vehicle fleet available* category, with the following elements:

- One vehicle

- <u>More than one vehicle</u>

This category may appear superfluous, but in terms of developing a modeling framework, formulating the mathematical program, and developing a solution algorithm, the distinction is important. SAMS operational problems are inherently *multiple* vehicle problems.

2.3 Taxonomic Categories to Classify SAMS Operational Problems

This section presents taxonomic categories in the literature that are still relevant to classify different SAMS operational problems. In some cases, this thesis slightly alters these taxonomic categories.

Vehicle Homogeneity

Bodin and Golden (1981) list the following taxonomic elements for their *type of vehicle fleet* category:

- Homogenous
- Heterogeneous

There are large differences in terms of formulating and solving homogenous and heterogeneous fleet operational problems. As there are likely to be SAMS providers with homogenous and heterogenous AV fleets, this category is still useful in terms of categorizing SAMS operational problems. This thesis renames the *type of vehicle fleet* category because it is too broad.

The AV fleets in the SAMS operational problems addressed in this thesis are homogenous.

Vehicle Capacity Constraints

Bodin and Golden (1981) list the following taxonomic elements for their *vehicle capacity constraints* category:

- Imposed all the time
- Imposed some of the time
- Not imposed

Once again this is an important consideration for modeling SAMS operational problems, as all three possibilities are still relevant. If the AVs are small vehicles and the mobility service options includes shared rides or ride-matching, then capacity constraints should be imposed all the time. However, if the AV fleet does not allow shared-rides, then capacity constraints are probably unnecessary. If the AVs are like buses or large vans, capacity constraints might be necessary in some cases and not others depending on overall demand levels.

This thesis examines problems with and without shared rides; hence, capacity constraints are (implicitly) imposed in the SAMS fleet operational problems with shared rides but are unnecessary in the problems without shared rides.

Maximum Vehicle Route Times and Distances

Bodin and Golden (1981) list the following taxonomic elements for their *maximum vehicle route times* category:

- Imposed – all the same
- Imposed – not all the same
- Not imposed

Maximum vehicle route times might be imposed all the time when the SAMS includes shared rides. The maximum vehicle route time parameter would prevent in-vehicle users from experiencing too long of a detour as their AV picks up and/or drops off other users. Lyft currently imposes such constraints for their Lyft Line service (Brown, 2016). Additionally, for non-AV mobility services, the maximum vehicle route time constraint could be used to model the maximum hours of service constraint on drivers. Additionally, as each driver has different preferences, the route time constraint could be dependent on the driver.

Regarding maximum vehicle route *distances*, this constraint is especially relevant when it comes to modeling electric AV fleets. Most electric vehicles have limited range compared to gasoline-powered vehicles and electric battery charging stations are sparse compared to gasoline refueling stations. In terms of modeling maximum vehicle route distances in a dynamic modeling framework, it is necessary to keep track of each AV’s fuel level.

In other SAMS operational problems, modeling maximum vehicle route time and distance constraints may not be necessary. Such instances include modeling the morning-peak period with a gasoline-powered AV fleet.

The *on-demand SAMS without shared rides* and the *on-demand autonomous carsharing service* problems in this study do not include maximum vehicle route time or distance constraints. However, the *on-demand shared-ride SAMS* problem does include a hard constraint on maximum route detour distance for individual travelers. In fact, sensitivity analysis results are presented as a function of the maximum detour distance for in-vehicle travelers.

Costs

Bodin and Golden (1981) list the following taxonomic elements in their *costs* category:

- Variable or routing costs
- Fixed operating or vehicle acquisition costs (capital costs)

In typical routing problems, variable operating costs are included in the objective function. These variable costs are typically a function of travel costs or travel distances (e.g. fuel costs and depreciation). Sometimes the problem formulation includes fixed vehicle acquisition costs in the objective function or a constraint on the fleet size. This classification category is still relevant for SAMS operational problems because modeling frameworks may incorporate variable costs, fixed costs, or both.

The objective functions for the SAMS operational problems in this thesis only include variable fleet costs.

Objective

Bodin and Golden (1981) list the following taxonomic elements in their *objectives* category:

- Minimize routing costs incurred
- Minimize sum of fixed and variable costs
- Minimize number of vehicles required

Savelsbergh and Sol (1995) also classify potential objective functions for the PDP; the elements include:

- Single vehicle objectives
    - Minimize duration
    - Minimize completion time

  - Minimize travel time
  - Minimize route length
  - Minimize client inconvenience
- Multiple vehicle objectives
  - Minimize number of vehicles
  - Maximize profit

Neither of these lists are exhaustive but Savelsbergh and Sol's (1995) list includes significantly more options. For the problem framework to incorporate profit maximization, the fleet operator must be able to reject customer requests or price discriminate. Section 2.4 discusses acceptance and rejection of requests as well as pricing options.

Per Savelsbergh and Sol (1995), objectives such as duration, completion time, and travel time have no clear meaning for dynamic problems. Hence, these objectives are not useful for SAMS operational problems. However, objectives such as minimizing cumulative travel time and/or wait time and minimizing cumulative vehicle miles traveled are appropriate in the dynamic context. Table 2-1 lists possible *objectives* for SAMS operational problems.

The objective used in the SAMS operational problems in this thesis is to minimize a combination of operational costs (i.e. fleet miles) and user inconvenience (i.e. wait times).

2.4 Taxonomic Categories Highly Relevant to SAMS Operational Problems

This section presents taxonomic categories that are highly relevant to SAMS operational problems. Moreover, they are not explicitly outlined in existing VRP taxonomies in the literature.

Fleet Size Elasticity

- Elastic
- Fixed fleet size

The sharing-economy and ridesourcing companies motivated this taxonomic category. Ridesourcing companies have highly elastic fleets, despite not owning any vehicles, because they can set transportation prices to attract drivers. If user demand exceeds vehicle supply, ridesourcing companies can increase prices to both decrease demand and increase supply. Zha et al. (2018) model elastic fleet sizes in there model of geometric matching and spatial pricing in ride-sourcing markets.

SAMS fleet managers may be able to increase fleet-size in the short-term by either paying for access to privately-owned AVs or allowing drivers with non-AVs to provide transportation service. This area is ripe for additional research.

The SAMS operational problems in this thesis assume a fixed fleet size.

Reservation Structure

- Short-term (slot-based) rentals
- Point-to-point service
- Mixed

Short-term AV rentals are similar to existing carsharing services except that instead of a user going to a designated carsharing parking spot to access a vehicle, the AVs can travel unoccupied to pick up user requests at their desired origin locations as well as reposition themselves throughout the service region. With short-term rentals, the user has complete control over the AV for a specified time-slot. Conversely, point-to-point service describes the service currently provided by ridesourcing companies, wherein, a user requests pickup and drop-off points and the AV transports the user between those two points. With short-term rentals, users can temporarily store items in the AV such as during a shopping trip. The mixed reservation option describes an AV fleet wherein the vehicles can either provide point-to-point service or be rented to users for time-slots. Pricing the mixed fleet option to take into consideration the opportunity costs of not having an AV available for the other reservation option is another open research area.

This thesis addresses operational problems associated with a short-term rental SAMS and a point-to-point SAMS.

Pricing

- No pricing
- Fixed pricing structure
- Pricing, with no fixed structure

Problems wherein the pricing structure is fixed do not allow the fleet operator to price discriminate across customers. Fixed pricing structures include user mileage- and/or travel time- based structures. In problems with no fixed pricing structure, the fleet controller can charge higher rates based on the time-of-day, the location of the customer's origin and or destination, or as a function of information about competing SAMS providers in the area. In competitive transportation markets the rates charged by carriers are usually modeled as a function of the trip's marginal cost and the price elasticity of demand facing the firm.

A large portion of freight transportation service is outsourced to commercial carriers and pricing structures vary across commercial trucking firms. A significant amount of research exists on pricing transportation service; however, a relatively small amount is integrated within a fleet management framework. Figliozzi et al. (2007) and Sayarshad and Chow (2015) present modeling frameworks that incorporate pricing strategies into fleet operational problems. Figliozzi et al. (2007) assume the carrier calculates the marginal cost of accepting an additional customer request to determine the price to charge customers.

When modeling AV fleet operational problems, not including pricing in the problem framework can still yield valuable results; however, researchers should consider pricing to increase the behavioral realism of the problem.

The SAMS operational problems addressed in this thesis do not incorporate pricing.

Accept/Reject Decision

- Unnecessary
- Fleet manager decision
- Customer decision

The accept or reject decision comes into play when the problem's objective function includes revenue, in addition to costs. In the case where revenue is not considered in the problem framework the acceptance/rejection decision is unnecessary.

If revenue and costs are both considered in the fleet management problem framework, but pricing is fixed, then the fleet manager must accept or reject customer requests as they arrive in real-time. If the marginal operational cost of serving the customer request exceeds the marginal revenue associated with the request, the fleet manager will likely reject the request. If parameter distributions are available, the fleet operator's decision to accept or reject should incorporate the opportunity cost of sending the vehicle to service the customer request. For example, in the *profitable VRP*, the decision problem involves determining the set of demands to serve in addition to how to group the demands into vehicles and schedule pickups/deliveries (Archetti et al., 2014).

In the case where revenue is considered in the objective function and the pricing structure is not fixed, the accept/reject decision lies in the hands of the customer. The fleet manager offers the user a price that the user can accept or reject. Currently, ridesourcing companies operate under this model wherein the SAMS provider offers a price to the customer through a smartphone application.

There is a long history of revenue management (also known as yield management) in transportation (McGill and van Ryzin, 1999) with applications in air transportation (Bertsimas and de Boer, 2005) and rail transport (Bilegan et al., 2015) that are relevant to pricing SAMSs as well as modeling the accept or reject decision.

The SAMS operational problems addressed in this thesis require the SAMS fleet operator to serve all user requests within the pre-defined geographical service region.

Reservation Timeframe (Degree of Dynamism)

- Immediate requests
- Minimum pre-reservation time
- Mixed

Immediate requests represent requests wherein customers want transportation service as soon as they request a ride (Psaraftis, 1980). Minimum pre-reservation time represents cases wherein customers must reserve transportation service a pre-defined period prior to the time they want to be serviced. In the mixed case, customers can request service immediately or pre-reserve service for a future time-period. As of early 2016, Uber and Lyft only allowed immediate requests; however, by 2017, both Uber and Lyft allowed users to make advanced requests in several cities.

From a fleet operations perspective and a firm profit-maximization perspective, allowing advanced demand requests can be both beneficial and disadvantageous depending on the circumstances. If demand is high relative to fleet size, and the fleet manager can charge high prices to customers making immediate requests, it is disadvantageous to have advanced demand requests for two reasons. First, presumably the advanced user requests receive 'locked-in' rates that are lower than the rates currently being charged by the fleet to immediate demand requests; therefore, the company loses money by serving the advanced user requests rather than the immediate requests. Second, the advanced requests add binding constraints to the fleet management problem when demand is high. Without the advanced requests, the vehicles would be free to focus on areas of high demand. Conversely, if demand is low relative to fleet size, it is beneficial to have advanced information on the location and time of demand requests to efficiently route vehicles and minimize empty vehicle miles.

This thesis exclusively addresses SAMS operational problems with immediate requests only. SAMS operational problems with immediate requests are referred to as *on-demand* SAMS operational problems. This has become the convention in the (autonomous) mobility service literature (Alonso-Mora et al., 2017; Pavone et al., 2012; Sayarshad and Chow, 2017).

Shared Rides

- No Shared Rides
- Shared Rides

Shared rides refer to the case where an AV can transport two or more different demand requests at the same time. Given a fixed-fleet size, the inclusion of shared rides typically reduces user wait time but increases user in-vehicle travel time. The well-known dynamic DARP includes shared rides (Psaraftis, 1980); however, most DARP problems were originally formulated within the context of transporting elderly and disabled users who, unfortunately, cannot operate vehicles. In contrast, SAMS providers are undoubtedly planning to offer service to all users, not just those unable to afford and/or operate a vehicle. To attract users that are wealthy and healthy enough to own and operate a vehicle, the service quality and convenience of an AV fleet will need to be commensurate with owning one's own vehicle. Hence, the increased importance of service quality and convenience in AV fleet operational problems, relative to the traditional DARP, will need to be considered in the modeling framework. Service quality and convenience can be included in the modeling framework via including user wait time and user travel time in the objective function. Additionally, hard constraints for wait time, travel time, and time-windows can be included in the mathematical formulation of the problem.

This thesis models two on-demand SAMSs without shared rides and an *on-demand shared-ride SAMS*. The next section provides a thorough review of stochastic dynamic VRPs (SDVRPs) without shared rides. However, given the modeling and algorithmic advances for the shared-ride SDVRP, it is important to refer readers to these studies. Most of shared-ride SDVRP research comes from freight transportation; however, Alonso-Mora et al. (2017) present a new method to solve on-demand shared-ride SDVRPs for urban passenger transportation. The method involves pre-computing person-to-person sharing opportunities and person group-to-vehicle sharing opportunities using share-ability networks (Santi et al., 2014), then solving an assignment problem between person groups and vehicles.

In freight transportation, several studies use Markov decision processes (MDPs) to model SDVRPs (Goodson et al., 2013; Thomas, 2007; Marlin Ulmer et al., 2017; Ulmer, 2017) and Ulmer et al. (2017) introduce route-based MDPs. Modeling SDVRPs as MDPs has become common in freight research, but less common in passenger SDVRPs. A freight SDVRP closely related to shared-use mobility service operational problems is the same-day delivery problem for online purchases (Voccia et al., 2017). Voccia et al. (2017) show that waiting at the depot for more users to request products can produce better solutions under certain circumstances. Other studies illustrate the benefits of holding/waiting strategies for SDVRPs (Pureza and Laporte, 2008; Thomas, 2007).

Repositioning of Vehicles

- No repositioning
- Repositioning

Repositioning or rebalancing AVs to subregions of the service region where supply is currently lower than expected future demand is a strategy to potentially increase the efficiency of an AV fleet. Repositioning strategies require stochastic information in the form of the spatio-temporal distribution of future traveler demand. For example, prior to the morning peak, repositioning AVs to residential areas is probably an effective strategy as demand is most likely to originate in residential areas. Similarly, before the afternoon peak, repositioning AVs into the CBD is probably an effective strategy. Models and algorithms for repositioning SAMSs can benefit from the extensive literature addressing ambulance/emergency vehicle repositioning and rebalancing problems (Andersson and Värbrand, 2007; Brotcorne et al., 2003; Gendreau et al., 2001; Nasrollahzadeh et al., 2018; Schmid, 2012; Schmid and Doerner, 2010).

This thesis addresses SAMS operational problems where stochastic information is unavailable, and repositioning is not considered, as well as problems where stochastic information is available and repositioning strategies are considered.

Underlying Network

- Real road network
- Test road network
- Graph or Virtual Network

In the classic VRP, the depot and customer origins or destinations comprise the nodes, and arcs connect each pair of nodes. This is referred to as a graph or virtual network, as it represents an abstraction of a real network. It also possible to incorporate physical road networks in the VRP where links represent streets and nodes represent street intersections. This taxonomy also distinguishes between real road networks and test road networks. Test networks do not represent real-world networks; rather, they are simplistic networks designed by an analyst such as grid networks.

This thesis models the underlying network for the SAMS operational problems as a uniform plane with Manhattan distances.

Network Congestion/Travel Times

- Static and Deterministic
- Time-dependent and Deterministic
- Time-dependent and Stochastic

In modeling frameworks with virtual or real networks, the network links may include congestion. This taxonomy mentioned previously that the congestion on links can be deterministic or stochastic. Additionally, congestion on links can fluctuate based on the time of the day (time-dependent) or remain constant throughout the analysis period (static).

In the models of SAMS operational problems, in this thesis, the AVs travel at a constant speed throughout the Manhattan Plane.

Demand Reassignments/Vehicle En-route Diversions

- Allowed
- Not Allowed

In most SDVRP modeling frameworks, after a vehicle is assigned to a demand request, this assignment decision is treated as irrevocable. Conversely, it is possible to allow vehicles to later be diverted while they are en-route to pick up demand requests. In this case, the demand must also be reassigned to another vehicle. Allowing demand reassignment and en-route vehicle diversions increases the potential solution space in SDVRPs; however, it often comes at the cost of larger computational times. It may also be infeasible (i.e. food delivery) or considered unprofessional (i.e. a driver expects driver A, but driver B shows up) to allow reassignments in some circumstances.

This thesis studies on-demand SAMS operational problems with and without en-route AV diversions and traveler reassignments.

Decision Epochs

- One at beginning (deterministic)
- Exogenous
- Endogenous

In SDVRPs, decisions need to be made as the system advances in time. The decision epochs can be determined endogenously or exogenously. In the exogenous case, the fleet operator decides when to re-optimize the system. It can be every 5 seconds, every 30 minutes, or anywhere in between. Conversely, the system can be modeled such that events in the system trigger a re-optimization of the system. In deterministic VRPs, the only decision period is prior to the start of vehicle routes.

The decision epochs in this study are determined exogenously.

2.5 Conclusion

This chapter presents a taxonomy of VRPs relevant to potential SAMS operational problems. Given the large number of taxonomic categories, there are numerous combinations of potential SAMS operational problems to study. This thesis focuses exclusively on *on-demand* SAMS operational problems. The next section reviews the literature most relevant to on-demand SAMS operational problems. It focuses mainly on SDVRPs without shared rides.

Chapter 3 Literature Review

3.1 Overview

This chapter reviews literature relevant to the real-time operation of SAMSs. On-demand SAMS operational problems are members of the class of stochastic dynamic vehicle routing problems (SDVRPs). After seminal work on the dynamic dial-a-ride problem (D-DARP) nearly forty years ago (Psaraftis, 1980), researchers have been developing models and solution algorithms for various SDVRP applications. SDVRP applications include taxi services, ambulance and other emergency services, paratransit services, and freight trucking services. According to a definition by Powell (1996), SDVRPs involve a vehicle fleet providing transportation service to demand requests that arrive dynamically and randomly, and require a fleet controller to assign vehicles to demand requests in real-time.

According to the classification presented by Berbeglia et al. (2010), there are three types of dynamic pickup and delivery problems (D-PDP), including the dynamic truckload pickup and deliver problem (D-TLPDP), the D-DARP, and the dynamic vehicle routing problem with pickup and delivery (D-VRPPD). Unlike the D-DARP and the D-VRPPD, the D-TLPDP only allows one demand request in a vehicle at a time. As this thesis analyzes SAMSs without shared rides, the D-TLPDP is the most relevant. However, the D-TLPDP typically refers to freight transportation applications in which service constraints are either non-existent or significantly less stringent than passenger transportation applications according. This literature review focuses on SDVRP applications without shared rides, including taxi dispatching problems, ambulance dispatching and/or repositioning problems, and freight D-TLPDPs. For readers interested in an in-depth review of SDVRPs, there are several recent reviews in the literature (Pillac et al., 2013; Psaraftis et al., 2016; Ritzinger et al., 2015). Readers may also be interested in a taxonomy and definition of rich VRPs (Lahyani et al., 2015).

3.2 Degree of Dynamism

Although all SDVRPs, by definition, are dynamic, the degree of dynamism varies considerably across applications. Lund et al. (1996) proposed a degree of dynamism (DoD) metric defined as the ratio of the number of dynamic demand requests (n_d) to the total number of requests (n_{tot}); ($DoD = n_d/n_{tot}$), wherein a dynamic request occurs while the vehicle fleet is providing service rather than before the start of the day. In the SAMS operational problems in this thesis, all the requests are dynamic (i.e. no requests are known before the start of the day). Hence, according to the DoD metric proposed by Lund et al., $DoD = 1$ for all on-demand SAMS operational problems.

The effective-DoD ($EDoD$) extends the DoD metric via considering the time individual demand requests become known to the operator, and the latest possible time the request could be received (Larsen et al., 2002). If t_r^i is the request time of demand i, and T is the length of the finite planning horizon, then $EDoD = \left(\sum_{i=1}^{n_d} \frac{t_r^i}{T}\right)/n_{tot}$. The $EDoD$ with time windows ($EDoDTW$)

considers the gap between user i's latest allowable departure time ($t_{d_l}^i$) and t_r^i (Larsen et al., 2002). The $EDoDTW$ is defined as follows, $EDoDTW = \frac{1}{n_{tot}} \sum_{i=1}^{n_d} \left(1 - \frac{t_{d_l}^i - t_r^i}{T}\right)$.

The on-demand SAMS operational problems in this thesis include soft, rather than hard, time window constraints; hence, l_i is not explicitly defined. However, $t_{d_l}^i - t_r^i$ is implicitly very small because the problem definitions assume users want to be picked up shortly after requesting a ride. Larsen et al. (2002) classify SDVRPs as weakly dynamic (e.g. distribution of gas and oil to households; the DARP for the elderly and physically disabled), moderately dynamic (e.g. overnight mail services; appliance repair), or strongly dynamic (e.g. police, fire, and ambulance services; taxicab services). The on-demand SAMS operational problems in this thesis can all be classified as strongly dynamic because in all the problems the users want to be picked up immediately after making a request and expect to be picked up within a few minutes.

Additionally, van Lon et al. (2016) present metrics that differentiate between dynamism and urgency. They define degree of dynamism as 'the continuity of change", where, in a very dynamic problem, new information (i.e. user request times) arrives continuously throughout the analysis period. On the other hand, the degree of urgency is an indicator of "the reaction time available for responding to an incoming [request]." The on-demand SAMS operational problems in this thesis have a high degree of dynamism according to the definition of van Lon et al. as user requests enter the system continuously throughout the period of analysis. Additionally, the on-demand SAMS operational problems (implicitly) have a high degree of urgency, as the fleet controller must assign AVs to new user requests very quickly and have AVs pick up user requests on the order of minutes. Unlike the $EDoD$ and the $EDoDTW$, the urgency and dynamism metrics in van Lon et al. (2016) apply to the infinite horizon case.

van Lon et al. (2016) highlight that measures of dynamism and urgency should be problem specific not dependent on algorithmic solution approach; however, measures of dynamism and urgency can assist in the choice of algorithmic approaches for specific problems.

3.3 SDVRPs Applications

This section reviews SDVRP applications relevant to on-demand SAMS operational problems without shared rides. These applications include taxi-dispatching problems, ambulance dispatching and fleet management problems, freight dynamic truckload pickup and delivery problems, and mobility-on-demand/autonomous mobility-on-demand problems.

Taxi dispatching

Before smartphones, users typically hailed taxicabs on the side of the road or called a taxi-dispatching service with their location. With taxi dispatchers, neither the user, the taxi driver, nor the taxi dispatcher had real-time information on the location of *both* the user and the taxi. Although, smartphones and real-time location information have changed the nature of personal transportation services (like taxis), research on taxi dispatching problems before smartphones is still relevant to the on-demand SAMS operational problems in this thesis.

Taxi dispatchers typically follow rule-based policies when dispatching taxis to users making phone requests. An early policy assigned each new phone request to the taxi that has been waiting the longest. Although equitable, the operational inefficiency of this policy is obvious, as taxis may have to travel a long way to pick up a request. A slightly less inefficient policy involves assigning each new phone request to the nearest idle taxi. However, research clearly shows that this highly-myopic policy is inefficient (Kiam Tian Seow et al., 2010). Chapter 6 re-illustrates the inefficiency of this strategy. Lee et al. (2004), in a simplified analysis, show that it is beneficial to assign new user requests to the taxi with the shortest network travel time, rather than the taxi with the shortest direct-line distance to the new user request. Unlike the SAMSs defined in this thesis, some taxi services and taxi problems defined in the literature allow taxi dispatchers to reject user requests (Kiam Tian Seow et al., 2010).

Maciejewski and Nagel (2013) present three different policies/strategies for the taxi-dispatching problem with immediate requests. The first strategy assigns users FCFS to the nearest idle taxi; the second and third strategy consider idle and en-route drop-off AVs in the assignment. The study explicitly considers the case where the demand rate of taxi requests temporarily outpaces the service rate of taxis. As a queue of unserved user requests forms, the fleet operator assigns multiple ordered requests to each taxi. In the third strategy, demand requests in each taxi's queue can be re-assigned to other taxis as new information enters the system.

Like the operational strategies in this thesis, Maciejewski et al. (2016) use the assignment (or bipartite matching) problem framework to dispatch taxis to immediate user requests. However, the taxi-traveler assignment strategies in Maciejewski et al. (2016) do not allow en-route pickup taxis (assigned users) to be diverted (reassigned), nor do they incorporate repositioning strategies.

Ambulance Dispatching and Fleet Management

The operational problems associated with ambulance and emergency services, like the on-demand SAMS operational problems defined in this thesis, have a high degree of urgency (van Lon et al., 2016). Given the nature of ambulance services, a few minutes difference in response times (i.e. patient waiting times) can have life-altering consequences. Hence, ambulance fleets focus heavily on reducing average response times as well as maximizing the percentage of patients with a response time less than a threshold value (e.g. four minutes for the U.S. National Fire Protection Association) (Nasrollahzadeh et al., 2018).

Gendreau et al. (2001) divide the ambulance fleet management problem into the ambulance-dispatching problem and the ambulance relocation problem. Early research focuses on the ambulance relocation problem while using simple heuristics for the ambulance-dispatching problem (Andersson and Värbrand, 2007; Gendreau et al., 2001). More-advanced methods use mathematical programming formulations that jointly consider the ambulance dispatch and relocation problem (Haghani and Yang, 2007). Other researchers employ approximate dynamic programming methods to solve the ambulance dispatching and relocation problem (Schmid, 2012). Lee (2012) examines the ambulance-dispatching problem in the presence of a disaster wherein the demand rate of calls outpaces the ambulance fleet's service rate.

Recently, Nasrollahzadeh et al. (2018) present ambulance operational policies that allow more flexibility in ambulance relocations. In addition to not automatically assigning the closest unoccupied ambulance to service requests (an inefficient policy), the study allows two different types of ambulance relocations. The first relocation type is ambulance redeployment in which ambulances do not automatically return to a preassigned base after serving a request, rather the ambulance can be redeployed to another base. The second relocation type is ambulance reallocation wherein ambulances can reactively or proactively relocate to other bases to increase coverage. In the proactive case, there may already be an idle ambulance at the base; however, if the expected demand is high enough, multiple ambulances at the base may be warranted.

Despite matching or exceeding the degree of dynamism and urgency associated with the on-demand SAMS operational problems in this thesis, there are constraints on the ambulance operational problem that on-demand SAMS fleet controllers do not encounter. Most importantly, SAMS user pickup and drop-off times are much shorter and more reliably short than ambulance on-site pickup or service times, and at-hospital drop-off times. Knowing SAMS user pickup and drop-off times are going to be short allows the SAMS controller to easily assign en-route drop-off AVs to new user requests; whereas, assigning a currently in-use ambulance to an open patient request is more challenging, riskier, and in some cases infeasible.

Freight Dynamic Truckload Pickup and Delivery

Although the dynamic freight truckload pickup and delivery problem (D-TLPDP) has a lower degree of dynamism and urgency than on-demand SAMS operational problems, much of the research on SDVRPs without shared rides comes from the D-TLPDP. Hence, the models and solution approaches for the D-TLPDP can inform research on on-demand SAMS operational problems without shared rides.

To solve the D-TLPDP, researchers commonly employ a rolling-horizon solution procedure that involves repeatedly re-solving a static mathematical programming problem (Fleischmann et al., 2004; Frantzeskakis and Powell, 1990; Yang et al., 2004, 1999). For the D-TLPDP, the two most-common mathematical programming problems solved in a rolling-horizon fashion are the TLPDP and the assignment (or bipartite matching) problem. The TLPDP model allows the fleet operator to sequence (or schedule) the pickup and delivery of multiple open demand requests, for each vehicle; whereas, the assignment problem matches each vehicle to at most one open demand request.

There is an important trade-off to consider when choosing between the assignment problem formulation and the TLPDP formulation. The first element of the trade-off is the computational efficiency of the problem formulations and the ability to solve the problems exactly or heuristically. The linear relaxation of the assignment problem, an integer programming problem, always returns integer solutions because the problem's constraint matrix is totally unimodular, thereby exact solutions to large problem instances can be obtained quickly. Conversely, the TLPDP formulation is NP-hard and exact solutions to moderate size problems cannot be obtained quickly. The second element of the trade-off are the benefits of sequencing multiple pickup and drop-offs. The assignment problem does not explicitly allow sequencing; whereas, the TLPDP does. In general, sequencing multiple pickups and deliveries is beneficial; however, its value depends on the type of problem. In freight transportation, where loads do not need immediate service (e.g. within 5-10 minutes of the request time), sequencing pickups and deliveries is highly beneficial. Conversely, in the case of an on-demand mobility service, users want to be picked up immediately; hence, placing a new user request in a sequence behind two or more other user requests means the new user is likely going to receive an unacceptably long wait time. Given the computational efficiency benefits of the assignment problem, and the lack of value of sequencing user requests for an on-demand mobility service, the SAMS operational strategies in this thesis utilize the assignment problem formulation, not the TLPDP formulation, in the rolling-horizon solution approaches.

An important finding in the freight D-TLPDP literature that this thesis replicates in the context of on-demand SAMSs is the operational efficiency benefits of allowing vehicle diversions (i.e. allowing users to be reassigned from one AV to another and allowing AVs to divert from picking up one traveler to go pick up another) in the toolbox of operational strategies (Regan et al., 1996, 1995). Practically speaking, switching a user between AVs is a lot simpler than switching a freight load between trucks in many cases. Hence, the potential benefits of en-route vehicle diversions/traveler reassignment in passenger transportation exceed the potential benefits in freight transport.

Mobility-on-demand (MOD) and autonomous-MOD (AMOD) Services

Over the past several years, researchers have begun to address mobility-on-demand (MOD) and autonomous-MOD (AMOD) operational problems. Additionally, several research groups are simulating shared-use autonomous vehicles (SAVs), mainly to understand their transportation planning and policy implications. This section provides a review of the relevant MOD, AMOD, SAV, and SAMS literature.

Simulating SAMS Fleets for Planning and Policy Implications

Existing supply-side research aiming to model SAMSs and understand their transportation planning and policy implications, generally involves an agent-based modeling framework with three components including a demand (i.e. user request) generator, an SAMS fleet controller, and some representation of the transportation network (Levin et al., 2017b; Rigole, 2014).

The demand generator creates user requests each with an origin location, destination location, and request time. Researchers have calibrated their spatio-temporal traveler demand generators using synthetic travel demand from regions such as Austin (Chen et al., 2016; Fagnant et al., 2015; Fagnant and Kockelman, 2016; Levin et al., 2017b), New Jersey (Zachariah et al., 2014; Zhu and Kornhauser, 2017), Lisbon (Martinez and Viegas, 2017; Viegas and Martinez, 2016), Berlin (Bischoff and Maciejewski, 2016), Munich (Dandl et al., 2017), and Zurich (Boesch et al., 2016). Other researchers use taxi data (Burns et al., 2013) or the National Household Travel Survey information (W. Zhang et al., 2015a) to obtain spatial-temporal demand distributions. This thesis generates synthetic demand, and uses taxi trip data from Manhattan, NY and Chicago, IL to represent demand.

In regards to the representation of the road network, seminal SAMS research models taxi stands but not the road network connecting the taxi stands (Ford, 2012; Zachariah et al., 2014). Other researchers use Manhattan grid networks (Fagnant and Kockelman, 2014) and Euclidean planes (Spieser et al., 2014) as abstractions of road networks. More advanced network representations include quasi-dynamic grid-based (W. Zhang et al., 2015b, 2015a) and quasi-dynamic actual road (Fagnant et al., 2015; Fagnant and Kockelman, 2016; International Transport Forum, 2015; Martinez and Viegas, 2017) networks with time-dependent, but deterministic link travel times. Recent research focusing on network impact assessment employs dynamic traffic simulation software such as MATSIM (Bischoff and Maciejewski, 2016) and a cell-transmission simulation model (Levin et al., 2017b) to model the AVs in a congestible road network. As this thesis focuses on the SAMS fleet dispatching problem, it employs a Manhattan grid network with fixed travel times.

The fleet controllers in this area of research use simplistic rules to assign AVs to users. Burns et al. (2013) assign users FCFS to the nearest idle or en-route drop-off AV. Zhang et al. (2015b) only consider idle AVs in their FCFS assignment strategy. Several researchers use a rule-based strategy that involves segmenting the service region into sub-regions, and assigning unassigned users (ordered randomly) to the closest idle AVs within their sub-region (Chen et al., 2016; Fagnant and Kockelman, 2014). If no idle AV is available in the user's sub-region, the user looks to the surrounding sub-regions. Boesch et al. (2016) employ a similar rule-based dispatching strategy. Fagnant et al. (2015) use a similar strategy in a road network setting and employ a modified-Dijkstra's algorithm to determine the shortest network path between idle AVs and unassigned users. Bischoff and Maciejewski (2016) model a large-scale SAMS and use a slightly more-sophisticated assignment strategy that involves classifying the system state into two mutually exclusive categories. If there is an oversupply of AVs relative to unassigned user requests, users are assigned FCFS to the nearest idle AV. If there is an undersupply of AVs, when an AV becomes idle it is assigned to the nearest unassigned user request. Chapter 6 employs optimization-based strategies to dynamically assign AVs to user requests, and additionally allows previously assigned users (en-route pickup AVs) to be reassigned (diverted) to other AVs (travelers) after new user requests enter the system.

MOD and AMOD Operational Problems

The smartphone, other ICT advancements, and real-time vehicle tracking devices motivated significant research on SDVRPs. Similarly, the emergence of MOD services and the promise of AVs and their inclusion in MOD fleets resulted in a large growth in SDVRP research for passenger transportation over the past five years (Alonso-Mora et al., 2017; Zhang and Pavone, 2016).

Several studies present model predictive control (MPC) approaches to control a fleet of AVs in a point-to-point AMOD service (Iglesias et al., 2017; Tsao et al., 2018; R. Zhang et al., 2015). A relevant difference between the problem definition (or model) of on-demand SAMSs in this thesis and in the MPC studies (Iglesias et al., 2017; Tsao et al., 2018; R. Zhang et al., 2015) relates to the treatment of origin and destination locations. The MPC studies only allow demand to originate and terminate at *stations* (representing regions of the service area); conversely, this thesis allows demand to originate and terminate anywhere in a Manhattan plane.

Existing point-to-point MOD and AMOD studies in the literature tend to focus on either the AV-user assignment decision problem (Alonso-Mora et al., 2017; Hyland and Mahmassani, 2018; Maciejewski et al., 2016; Maciejewski and Nagel, 2013) or the empty AV repositioning decision problem (Iglesias et al., 2017; Sayarshad and Chow, 2017; Winter et al., 2017; Zhang et al., 2018), rather than a detailed treatment of both decision problems. Conversely, Chapter 8 treats these two decision problems equally (or rather jointly) as they are interdependent, and both impact fleet performance.

In a comparison of point-to-point AMOD operational strategies, Horl et al. (2017) compare assignment-only strategies, such as an AV-user assignment heuristic and a bipartite matching of available AVs and open user requests, against strategies that combine bipartite matching for the AV-traveler assignment decision problem with the AV repositioning approach in Pavone et al. (2011). However, the operational strategies in Horl et al. (2017) solve the two decision problems sequentially, rather than jointly.

Shared Rides

The operational problem associated with shared-ride service is similar to the D-DARP that was first introduced 40 years ago (Psaraftis, 1983, 1980; Wilson and Colvin, 1977). The original D-DARP was formulated in the context of paratransit service for individuals with disabilities. The shared-ride problem presented in this thesis is significantly more dynamic than the initial D-DARP; moreover, the level-of-service associated with paratransit and other traditional dial-a-ride services are inferior to the level-of-service provided by existing ridesourcing and necessary to compete with the personal (autonomous) vehicle in terms of service quality (i.e. wait time).

Recently, researchers have formulated and presented solution methods for the shared-ride taxi problem (Hosni et al., 2014; Santos and Xavier, 2015). Unlike the current thesis, these papers assume explicit time-window constraints on the traveler pickup and drop-off times. In the most-closely related research, Ma et al. model a taxi fleet that allows shared rides (Ma et al., 2015). They claim that taxi-sharing increases the number of travelers served by a factor of three, relative to taxi service without shared rides. Additionally, Martinez and Viegas use an agent-based simulation to model a shared, self-driving taxi service (Martinez and Viegas, 2017). Alonso-Mora et al. present an efficient and scalable algorithm to facilitate shared-ride service for a large fleet of taxis (Alonso-Mora et al., 2017).

Two recent survey papers review dynamic ridesharing research (Agatz et al., 2012; Furuhata et al., 2013). Ridesharing and dynamic ridesharing historically referred to a carpool-like service wherein a driver notifies the system operator of a trip he is planning to take (notification must include origin location, destination location, as well as the planned origin departure time and/or planned destination arrival time). Similarly, travelers notify the system operator of trips they plan to take. The system operator then (attempts to) match driver and traveler requests. This is the ridesharing service reviewed in (Agatz et al., 2012; Furuhata et al., 2013). Lee and Savelsburgh (2015) examine the case of a conventional ridesharing system complemented by a fleet of dedicated drivers.

Several existing SAMS studies include shared-ride service, which they refer to as dynamic ridesharing (Fagnant and Kockelman, 2016; Levin et al., 2017b; Rigole, 2014; W. Zhang et al., 2015a). This thesis refrains from referring to an (on-demand) shared-ride SAMS as (dynamic) ridesharing to prevent confusion among readers who associate ridesharing with a carpool-like service that does not include dedicated drivers or dedicated service vehicles.

3.4 Conclusion

This chapter reviews the SDVRP literature that is most related to on-demand SAMSs. The relevant SDVRP applications in the literature include taxi-dispatching problems, ambulance dispatching and fleet management problems, and freight dynamic pickup and delivery problems. The chapter also reviews the existing literature examining on-demand SAMSs. While on-demand SAMSs share many attributes with problems in the existing dynamic freight routing literature, taxi-dispatching literature, and ambulance-dispatching literature, the combination of the SAMS operational problems' size, degree of dynamism, degree of urgency, spatial distribution of user requests, and short user pickup and drop-off times make the problem instances unique relative to the existing literature.

Chapter 4 Operational Efficiency of Chicago Taxi Fleet[4]

4.1 Overview

This chapter evaluates and quantifies the operational efficiency of the Chicago taxi fleet using a spatial efficiency metric and a temporal efficiency metric. The spatial efficiency metric determines the percentage of a taxi's total miles that are productive. Productive miles are defined as miles wherein the taxi is transporting users and generating revenue. Unproductive miles include miles spent roaming for users and miles spent traveling to pick up users. The temporal efficiency metric determines the percentage of in-service time taxis are productive. In-service time includes all the time a taxi is looking for users or serving users; i.e. all the time a driver is in the taxi. The temporal and spatial efficiency measures allow the analyst, to assess the efficiency of individual taxi trips, individual taxis, and the entire taxi fleet. To characterize and quantify the efficiency of taxis, this chapter first clusters individual taxis based on their daily usage rates over the entire year.

This chapter utilizes taxi data from the city of Chicago (Chicago Data Portal, 2017) and serves two purposes in the context of this thesis. First, the spatial efficiency of the Chicago taxi fleet provides a baseline to analyze the efficiency of the SAMS operational policies presented and tested later in this thesis. Second, the results in this chapter show the inefficiencies associated with

[4] This chapter parallels the taxi fleet efficiency component of Chen and Hyland et al. (2018)

decentralized operation of a vehicle fleet, allowing drivers to completely control their own vehicles, and not having full information about open user requests.

4.2 Chicago Taxi Trip Data

Chicago followed in the footsteps of New York City who released 2009 through 2015 taxi data that covers 1.1 billion taxi trips. The New York dataset has spurred a significant volume of research (Haggag et al., 2017; King and Saldarriaga, 2017; Liu et al., 2015; Qian and Ukkusuri, 2015; Schneider, 2016; Yang and Gonzales, 2017; Zhan et al., 2014). Schneider (2016) presents a comprehensive exploratory analysis and visualization of the New York City taxi data. Zhan et al. (2016, 2014) extensively study the efficiency of urban taxi fleets and present a graph-based approach to evaluate the efficiency of a taxi fleet.

The taxi trip data used in this analysis is a subset of the taxi dataset available through the Chicago Data Portal (Chicago Data Portal, 2017). The dataset includes information on over 105 million taxi trips made in the city of Chicago between January 2013 and 2017. The full dataset includes a total of 23 columns with information about each taxi trip. This study makes use of the following columns:

- Taxi ID
- Trip Start Timestamp (15-minute interval)
- Trip End Timestamp (15-minute interval)
- Trip Duration (seconds)

- Trip Distance (miles)
- Pickup Location (census tract centroid or community area centroid)
- Drop-off Location (census tract centroid or community area centroid)

Due to privacy concerns, the dataset does not provide exact geographical coordinates for pickup or drop-off locations; rather it provides census tract or community area centroids.

Table 4-1 displays the number of taxi trips recorded, and the number of unique taxis that made at least one trip in each year between 2013 and 2016. Table 4-2 displays the distribution of taxi trip distances, durations, and trip counts.

In this analysis, taxi trips with distances (durations) longer than 40 miles (90 minutes) were either removed from each analysis, replaced with the shortest path distance (time) between the trip's pickup and drop-off locations using the Google Maps API, or simply replaced with a distance (time) of 40 miles (90 minutes). The 40-mile cut-off was selected because this is approximately the longest possible travel distance between any two locations in the Chicago Metropolitan region. Similarly, the 90-minute cut-off was selected because it is approximately the travel time from the northern-most edge to the southern-most edge of the Chicago metropolitan region, during the off-peak period.

Table 4-1: Chicago Taxi Trip Dataset General Statistics

	2013	2014	2015	2016
Number of Taxis	5,557	7,582	7,552	7,667
Number of Trips	26,870,287	31,021,726	27,400744	19,878,249

Table 4-2: Individual Taxi Trip Statistics

	Distance (mile)	Duration (minute)	Trip Count per Taxi (2013~2016)
Min.	0.11	1	1
1^{st} Quartile	0.9	6	429
Median	1.7	10	14,210
Mean	3.8	14	13,150
3^{rd} Quartile	3.9	17	22,160
Max.	40	90	48,020

4.3 Taxi Fleet Usage Analysis

Histogram of Taxi Fleet by Average Daily Trips

This section presents a histogram of the average number of trips taxis made per in-service day. Let J_y denote the set of all taxis, in year y, indexed by taxi $j \in J_y$. Let $N_{j,y}$ denote the set of trips completed by taxi $j \in J_y$ in year y, indexed by trip $n \in N_{j,y}$. For notational simplicity, and the fact there are separate analysis for each year y, the subscript y is removed from these sets throughout the remainder of this chapter. Equation (1) determines the average number of daily trips for taxi j (τ_j) in year y, where σ_j is the number of days taxi j is in service during year y.

$$\tau_j = \frac{|N_j|}{\sigma_j} \tag{1}$$

Figure 4-1 displays a histogram of τ_j for each year y. In 2015 and 2016, around 1,500 taxis only made three trips per *in-service* day on average. However, most taxis completed between 7 and 25 trips per in-service day, with a few taxis completing more than 40 trips per in-service day.

This histogram clearly does not provide the whole story in terms of usage of taxi vehicles. For example, it does not differentiate between vehicles based on work hours (some cabs may only be in use a couple hours per day) or work days (some cabs may only work Sundays). However, the histogram does suggest that there were many drivers (more so in 2015 and 2016) who only serve a couple of travelers per in-service day. This finding informs other results presented later in this chapter.

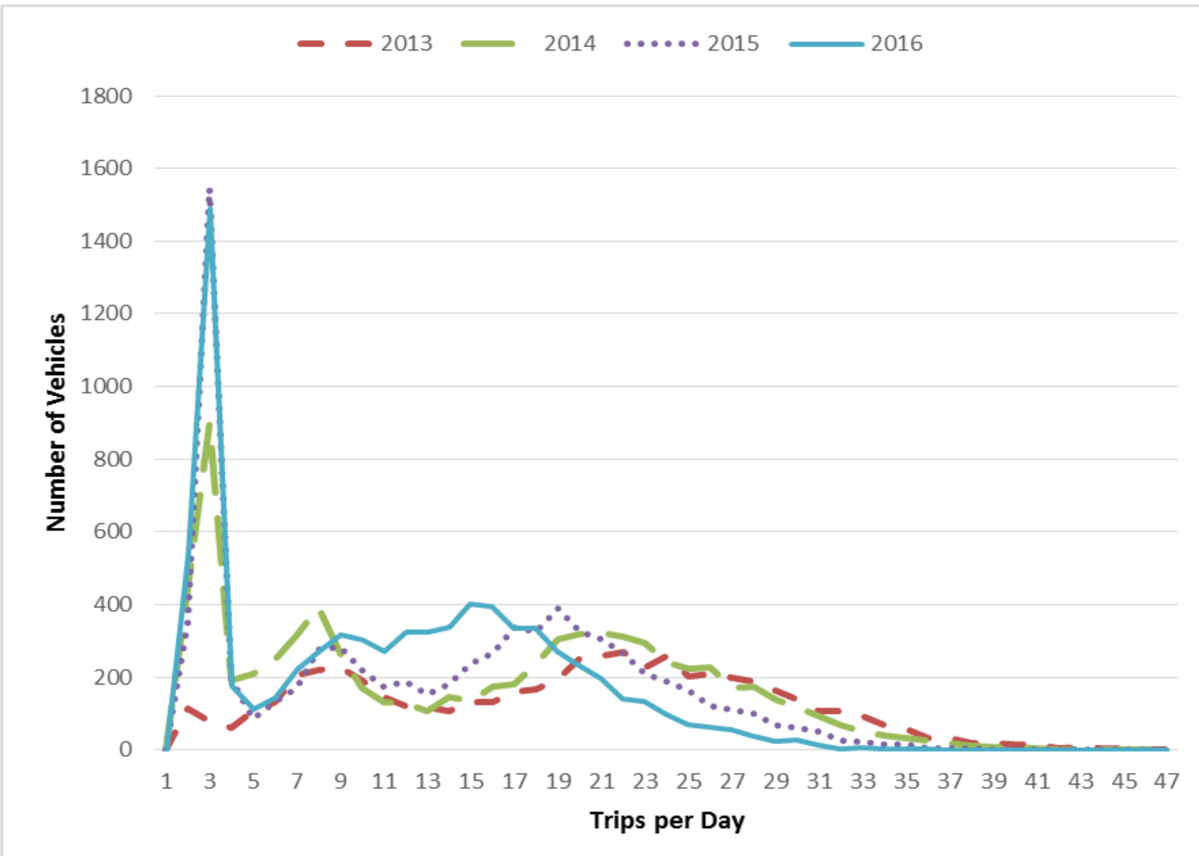


Figure 4-1: Distribution of the average number of daily trips (per in-service day) across taxis

Clustering Taxis by Daily Trip Count

This section describes the K-means clustering algorithm employed to cluster taxis based on their daily trip counts. Let D be the set of days in a year, indexed by $d \in D$, wherein $|D|$ is typically

365 (or 366 in a leap year). Let γ_{jd} denote the number of trips made by taxi $j \in J$ on day $d \in D$, and let Γ_j denote the taxi daily trip count vector of length $|D|$ for taxi j ($\gamma_{jd} \in \Gamma_j$).

The K-means clustering problem involves assigning each taxi $j \in J$ into one and only one cluster. Let S denote the set of clusters, where $|S| = K$. The elements of S are denoted $S_k,\ k = 1,2,\dots,K$, where S_k is a set of taxis ($S_k \subset J$). Let $\pi_{S_k,d}$ denote the mean number of trips for the taxis in cluster S_k on day d, and Π_{S_k} denote the mean daily trip count vector for the taxis assigned to cluster S_k ($\pi_{S_k,d} \in \Pi_{S_k}$), where $\left|\Pi_{S_k}\right| = \left|\Gamma_j\right| = |D|$. Equation (2) displays the squared error between the mean daily trip values for cluster S_k ($\pi_{S_k,d}$) and the daily trip values for each taxi j assigned to cluster S_k ($\gamma_{jd} |\, j \in S_k$).

$$SE(S_k) = \sum_{j \in S_k} \left\| \Gamma_j - \Pi_{S_k} \right\|^2 = \sum_{j \in S_k} \sum_{d=1}^{|D|} \left(\gamma_{jd} - \pi_{S_k,d} \right)^2 \tag{2}$$

A k-means clustering algorithm identifies the set of clusters (S) that minimize the within-cluster sum of squares (WCSS) as defined in Eqn. (3).

$$W(K) = \min_S \sum_{k=1}^{K} SE(S_k) = \min_S \sum_{k=1}^{K} \sum_{j \in S_k} \left\| \Gamma_j - \Pi_{S_k} \right\|^2 \tag{3}$$

where, K is the number of clusters and $W(K)$ is the minimal WCSS. $W(K)$ is a monotonically decreasing function of K. Equation (4) displays the metric used in this chapter to help select the number of clusters (K), where R is the rate of change (decrease) in $W(K)$ as K increases.

$$R = \frac{W(K) - W(K+1)}{W(K)} \times 100\% \tag{4}$$

For the Chicago taxi data, Eqn. (4) yields five taxi clusters. Table 4-3 shows the percentage of taxis in each of the five taxi-clusters by year.

Table 4-3: Percentage of Taxis in Each Cluster

Cluster	2013	2014	2015	2016
Most trips	17.06%	15.18%	8.47%	7.54%
More trips	11.77%	24.48%	20.19%	17.54%
Normal	25.52%	7.24%	20.52%	18.49%
Less trips	22.10%	20.76%	18.78%	17.67%
Least trips	23.56%	32.34%	32.03%	38.75%

Figure 4-2 presents the daily taxi-cluster centroids for each year, where the x-axis displays d and the y-axis displays $\pi_{S_k,d}$. In 2013, the range of average daily trip counts for taxis making the *most trips* was between 20 and 35; the range for taxis making *more trips* was between 15 and 25; the range for taxis with *normal trips* was 1-to-2 for the first 90 days, before increasing steadily over the next 150 days, and then plateauing between 15 and 25 for the last part of the year; the range for taxis making *less trips* was between 5 and 10; and the range for taxis making the *least trips* was one or two for the first 250 days, before it increased to between 5 and 10. The *least trips* and *less trips* clusters likely correspond to the three trips per in-service day peak in Figure 4-1. The 2014 daily trips ranges are like the 2013 daily trip ranges for all five clusters.

In 2015, the daily trip counts in each cluster were stable; i.e. no systematic increases or decreases in daily trip counts after a certain day of the year. The same was largely true for 2016, except the *less trips* cluster range decreased from 5-10 daily trips the first 200 days of the year to 1-3 daily trips during the last 165 days.

The results in Figure 4-2 clearly indicate that there are significant differences in taxi trips per day in the Chicago taxi fleet. Some taxis consistently made 20-35 trips per day and others only made one-two trips per day.

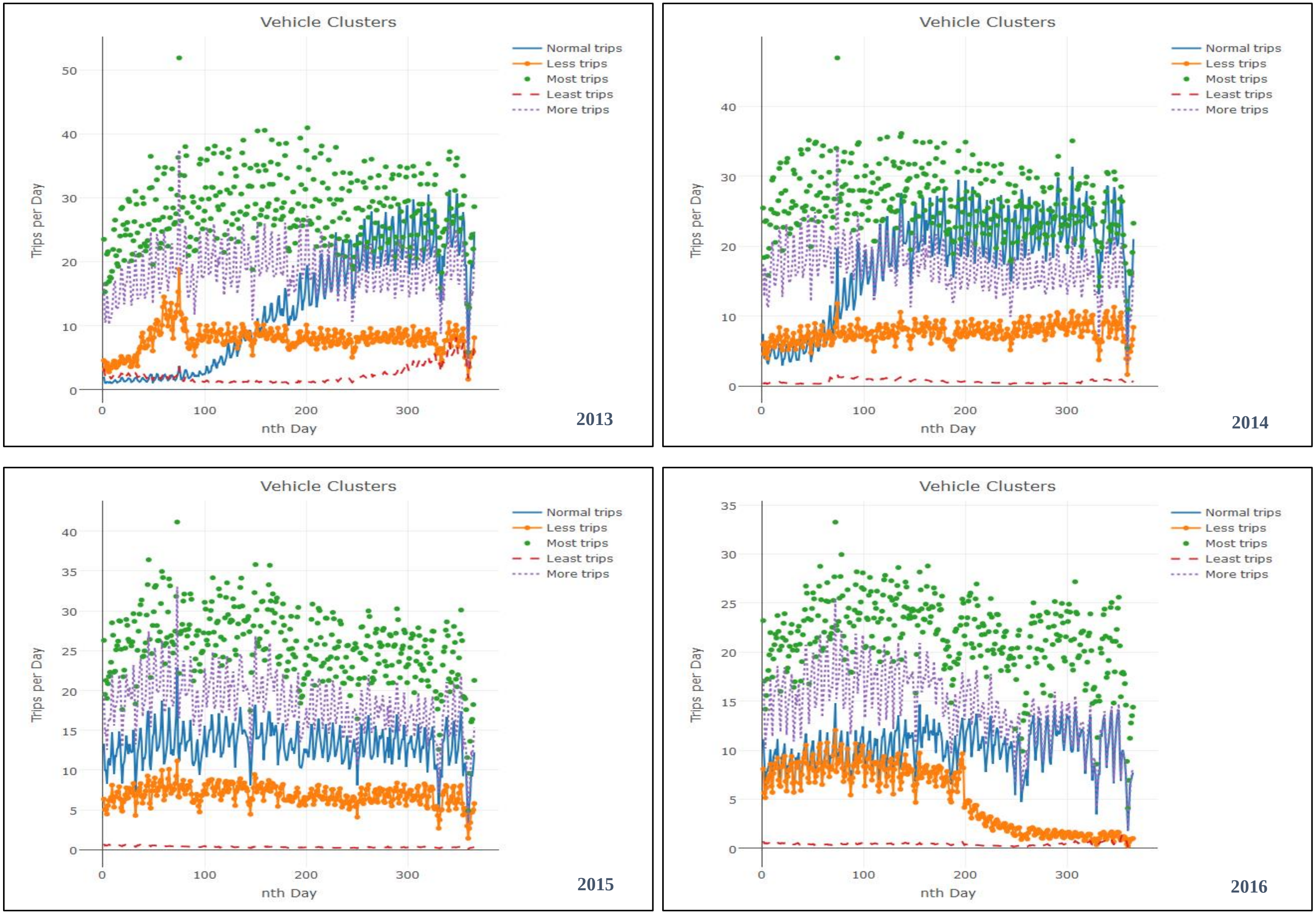


Figure 4-2: Taxi clusters results

Next, this section analyzes the percentage of in-service days for the taxis in each cluster to validate the results in Figure 4-2. Figure 4-3 presents histograms of the percentage of in-service days for the taxis in each cluster, independent of the average number of fares/trips per day.

In Figure 4-3, taxis in the *most trips* cluster were in-service 70-100% of the days in each year. Similarly, as expected, most taxis in the *least trips* cluster were only in service 0-40% of the days each year. The other graphs in Figure 4-3 also largely follow what one would expect based on the results in Figure 4-2. One exception is that the many taxis in the *less trips* cluster were in-service a large percentage of the days. Based on Figure 4-2, they must have served only a few trips per day.

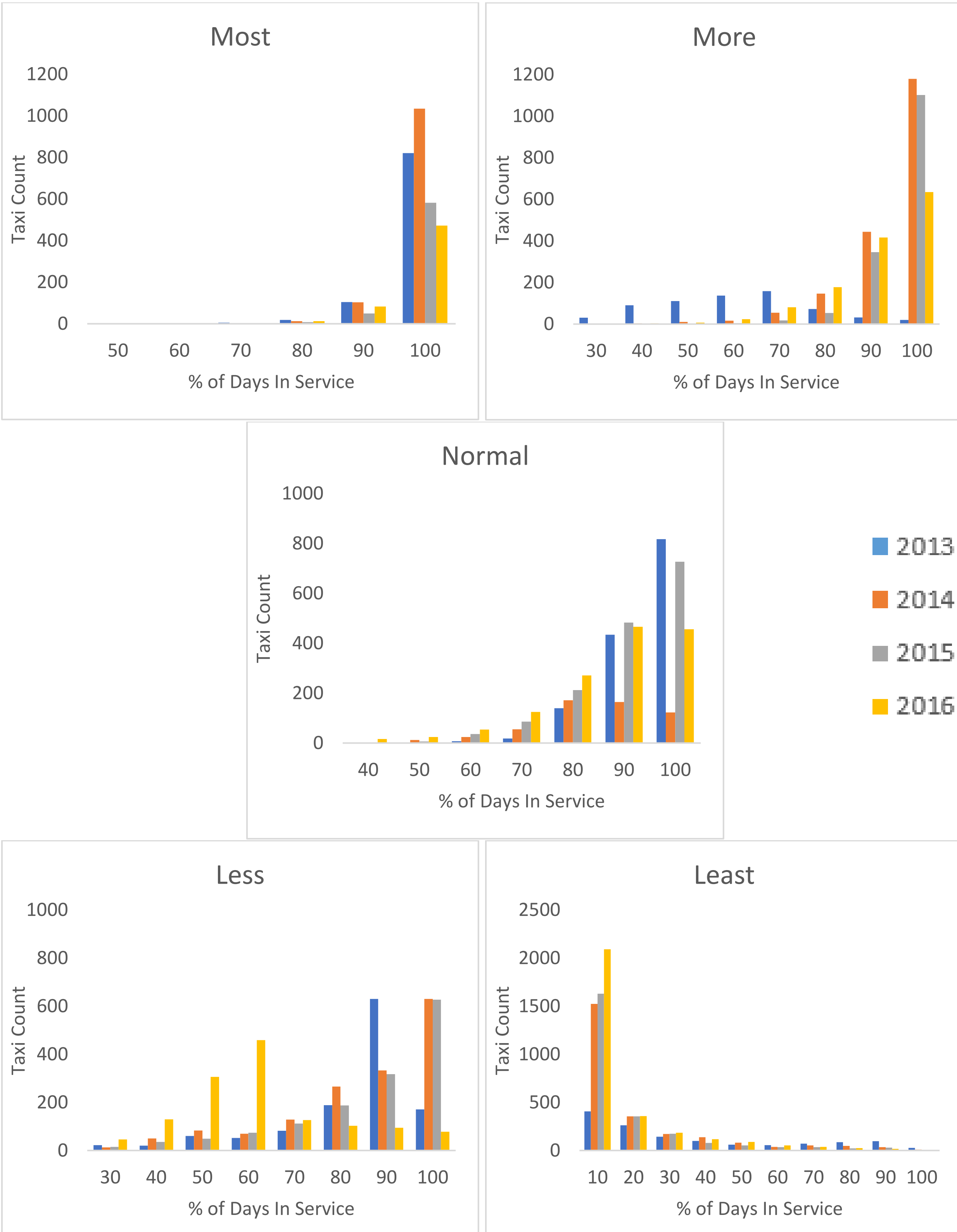


Figure 4-3: Histograms for the percentage of days in service for each taxi cluster

4.4 Taxi Efficiency Analysis

The last section focused on the usage rate of individual taxis, whereas this section focuses on the efficiency of individual taxis.

Efficiency Metrics

Temporal Efficiency Metric

Let $p_{j,n}$ denote the productive time associated with taxi-trip j, n, and let and $u_{j,n}$ denote the unproductive time between taxi-trip j, n and taxi-trip $j, n + 1$. Let $dura_{j,n}$ denote the trip duration value in the taxi dataset and let $b_{j,n}$ denote a binary variable equal to 1 if a trip duration value exists for taxi-trip j, n. Let $pos_{j,n,p}$ and $pos_{j,n,d}$ denote the pickup and drop-off locations (longitude and latitude) of taxi-trip j, n, respectively. Finally, let $A_{j,n,p}$ and $A_{j,n,d}$ denote binary variables equal to 1 if the pickup and drop-off locations of taxi-trip j, n exist in the dataset, respectively. Algorithm 1a assigns values to $p_{j,n}$ (the Google API function is described below).

Algorithm 1a

```
for all j ∈ J:     //All taxis
    for all n ∈ N_j:     //All taxi-trips
        if b_{j,n} == 1 and dura_{j,n} < 90min:     //If trip duration exists and is less than 90 minutes
            p_{j,n} = dura_{j,n}     // Assign trip duration value from taxi dataset
        else if b_{j,n} == 0 and (A_{j,n,p} == 0 or A_{j,n,d} == 0):     // If trip duration and pickup or drop-off
                                                                      //location do not exist in dataset
            Remove j,n from temporal efficiency analysis   //Necessary data not available
        else:
            if pos_{j,n,p} == pos_{j,n,d}:  //If n's pickup and drop-off census tract are the same
                Remove j,n from temporal efficiency analysis   //Same census tract, duration equals zero
            else:
                temp1 = GoogleAPIdura(pos_{j,n,p}, pos_{j,n,d})   //Call Google API
                p_{j,n} = min(90min, temp1)     //Assign minimum of API call or 90 minutes
```

Figure 4-4a: Algorithm to obtain the productive time of Taxi-trip j, n.

Similarly, Algorithm 1b assigns values to $u_{j,n}$. Let $t_{j,n,p}$ and $t_{j,n,d}$ denote the trip pickup and drop-off time for taxi-trip combination j, n, respectively. Moreover, let $B_{j,n,p}$ and $B_{j,n,d}$ denote binary variables equal to 1 if these values exist in the dataset for pickups and drop-offs, respectively.

Algorithm 1b

for all $\boldsymbol{j} \in \boldsymbol{J}$**:** //All taxis

  for all $\boldsymbol{n} \in \boldsymbol{N_j}$**:** //All taxi-trips

    if $\boldsymbol{B_{j,n,d} == 1 \; and \; B_{j,n+1,p} == 1}$**:** //If drop-off time of trip n and pickup time of trip $n+1$ exist

      $\Delta\boldsymbol{t} = \left(\boldsymbol{t_{j,n+1,p} - t_{j,n,d}}\right)$ //Time gap between drop-off time of trip n and pickup time of trip $n+1$

      if $\Delta\boldsymbol{t} < \boldsymbol{90min}$**:** //If time gap less than 90 minutes

        $\boldsymbol{u_{j,n}} = \Delta\boldsymbol{t}$ //Assign time gap as unproductive+ time

      else if $\Delta\boldsymbol{t} > \boldsymbol{120min}$**:** //If time gap greater than 120 minutes

        Remove $\boldsymbol{j, n}$ **from temporal efficiency analysis** //Driver is on a break or went home

      else:

        if $\boldsymbol{pos_{j,n,d} == pos_{j,n+1,p}}$**:** //If n's drop-off census tract the same as $n+1$'s pickup tract

          Remove $\boldsymbol{j, n}$ **from temporal efficiency analysis** //Same census tract, duration is zero

        else:

          $\boldsymbol{temp2} = \boldsymbol{GoogleAPIdura}\left(\boldsymbol{pos_{j,n,d}, pos_{j,n+1,p}}\right)$ //Call Google API

          $\boldsymbol{u_{j,n}} = \mathbf{min}(\boldsymbol{90min},\ \boldsymbol{temp2})$ //Assign minimum of API call or 90 minutes

    else if $\boldsymbol{A_{j,n,d} == 1 \; and \; A_{j,n+1,p} == 1}$ **:** // If drop-off location of n and pickup location of $n+1$ exist

      if $\boldsymbol{pos_{j,n,d} == pos_{j,n+1,p}}$**:** //If n's drop-off census tract the same as $n+1$'s pickup census tract

        Remove $\boldsymbol{j, n}$ **from temporal efficiency analysis** //Same census tract, duration equals zero

      else:

        $\boldsymbol{temp3} = \boldsymbol{GoogleAPIdura}\left(\boldsymbol{pos_{j,n,d}, pos_{j,n+1,p}}\right)$ //Call Google API

        $\boldsymbol{u_{j,n}} = \mathbf{min}(\boldsymbol{90min},\ \boldsymbol{temp3})$ //Assign minimum of API call or 90 minutes

    else:

      Remove $\boldsymbol{j, n}$ **from temporal efficiency analysis** //Necessary data not available

Figure 4-4b: Algorithm to obtain the unproductive time between taxi-trip $\boldsymbol{j, n}$ and taxi-trip $\boldsymbol{j, n+1}$

Equation (5) displays the temporal efficiency metric for taxi-trip j, n ($\varphi_{j,n}$) as a function of $p_{j,n}$ and $u_{j,n}$ obtained via Algorithm 1a and Algorithm 1b, respectively. Values close to 100% represent temporally efficient taxi trips; whereas, values close to 0% represented temporally inefficient taxi trips.

$$\varphi_{j,n} = \left(\frac{p_{j,n}}{u_{j,n} + p_{j,n}}\right) \times 100\% \tag{5}$$

Spatial Efficiency Metric

Let $l_{j,n}$ denote the loaded distance associated with taxi-trip j, n. Similarly, let $e_{j,n}$ denote the empty distance between the drop-off location of taxi-trip j, n and the pickup location of taxi-trip $j, n+1$. Let $dist_{j,n}$ denote the trip distance value in the taxi dataset and let $a_{j,n}$ denote a binary variable equal to 1 if a trip distance value exists for taxi-trip j, n. Algorithm 2a and Algorithm 2b assign values to $l_{j,n}$ and $e_{j,n}$, respectively.

Algorithm 2a

---

for all $\boldsymbol{j} \in \boldsymbol{J}$**:** //All taxis

  for all $\boldsymbol{n} \in \boldsymbol{N_j}$**:** //All taxi-trips

    if $\boldsymbol{a_{j,n} == 1}$ **and** $\boldsymbol{dist_{j,n} < 40\ miles}$ **:** //If trip distance exists and is less than 40 miles

      $\boldsymbol{l_{j,n} = dist_{j,n}}$ // Assign trip distance value from dataset

    else if $\boldsymbol{a_{j,n} == 0}$ **and** $\boldsymbol{(A_{j,n,p} == 0\ or\ A_{j,n,d} == 0)}$: // If trip distance and pickup or drop-off //location do not exist in dataset

      Remove $\boldsymbol{j, n}$ **from spatial efficiency analysis** //Necessary data not available

    else:

      if $\boldsymbol{pos_{j,n,p} == pos_{j,n,d}}$**:** //If n's pickup and drop-off census tract are the same

        Remove $\boldsymbol{j, n}$ **from spatial efficiency analysis** //Same census tract, distance equals zero

      else:

        $\boldsymbol{temp4 = GoogleAPIdist(pos_{j,n,p}, pos_{j,n,d})}$ //Call Google API

        $\boldsymbol{l_{j,n} = \min(40\ miles,\ temp4)}$ //Assign minimum of API call or 40 miles

---

Figure 4-5a: Algorithm to obtain the loaded distance of taxi-trip $\boldsymbol{j, n}$

Algorithm 2b

for all $\boldsymbol{j} \in \boldsymbol{J}$**:** //All taxis

for all $\boldsymbol{n} \in \boldsymbol{N_j}$**:** //All taxi-trips

$\Delta \boldsymbol{t} = (\boldsymbol{t_{j,n+1,p}} - \boldsymbol{t_{j,n,d}})$ //Time gap between drop-off time of trip n and pickup time of trip $n+1$

if $\boldsymbol{A_{j,n,d}} == \boldsymbol{0}$ $\boldsymbol{and}$ $\boldsymbol{A_{j,n+1,p}} == \boldsymbol{0}$ //If n's drop-off location and $n+1$'s pickup location do not exist

Remove $\boldsymbol{j}, \boldsymbol{n}$ **from spatial efficiency analysis** //Necessary data not available

else if $\Delta \boldsymbol{t} > \boldsymbol{120 min}$**:** //If time gap greater than 120 minutes

Remove $\boldsymbol{j}, \boldsymbol{n}$ **from spatial efficiency analysis** //Driver is on a break or went home

else:

if $\boldsymbol{pos_{j,n,d}} == \boldsymbol{pos_{j,n+1,p}}$**:** //If n's drop-off census tract the same as $n+1$'s pickup census tract

Remove $\boldsymbol{j}, \boldsymbol{n}$ **from spatial efficiency analysis** //Same census tract, distance equals zero

else:

$\boldsymbol{temp5} = \boldsymbol{GoogleAPIdist}(\boldsymbol{pos_{j,n,d}}, \boldsymbol{pos_{j,n+1,p}})$ //Call Google API

$\boldsymbol{e_{j,n}} = \mathbf{min}(\mathbf{40}\ \boldsymbol{miles},\ \boldsymbol{temp5})$ //Assign minimum of API call or 40 miles

Figure 4-5b: Algorithm to obtain the empty distance between taxi-trip $\boldsymbol{j}, \boldsymbol{n}$ and taxi-trip $\boldsymbol{j}, \boldsymbol{n+1}$.

Equation (6) displays the spatial efficiency metric for taxi-trip j, n ($\xi_{j,n}$) as a function of $l_{j,n}$ and $e_{j,n}$ obtained via Algorithm 2a and Algorithm 2b, respectively. Values close to 100% represent spatially efficient taxi trips; whereas, values close to 0% represented spatially inefficient taxi trips.

$$\xi_{j,n} = \left(\frac{l_{j,n}}{e_{j,n} + l_{j,n}}\right) \times 100\% \tag{6}$$

Google Maps Distance Matrix API

The Google Maps Distance Matrix API[5] is a service provided by Google Inc. that estimates travel distance and time for a recommend route between origin and destination points (Google, 2017). API requests must include origin and destination locations, a unique Google API key, and a transport mode. Optional parameters include arrival or departure time, and traffic model (best

[5] https://developers.google.com/maps/

guess, pessimistic, optimistic). This thesis uses the default driving mode and does not provide specific arrival or departure times. The API returns trip distance and duration for any feasible origin-destination coordinate pair input.

The analysis involves calculating the shortest travel time and distance between each unique pickup and drop-off location pair. Hence, the algorithms presented in the last two subsections use lookup table values for API-generated trip distances and trip durations.

Efficiency Results

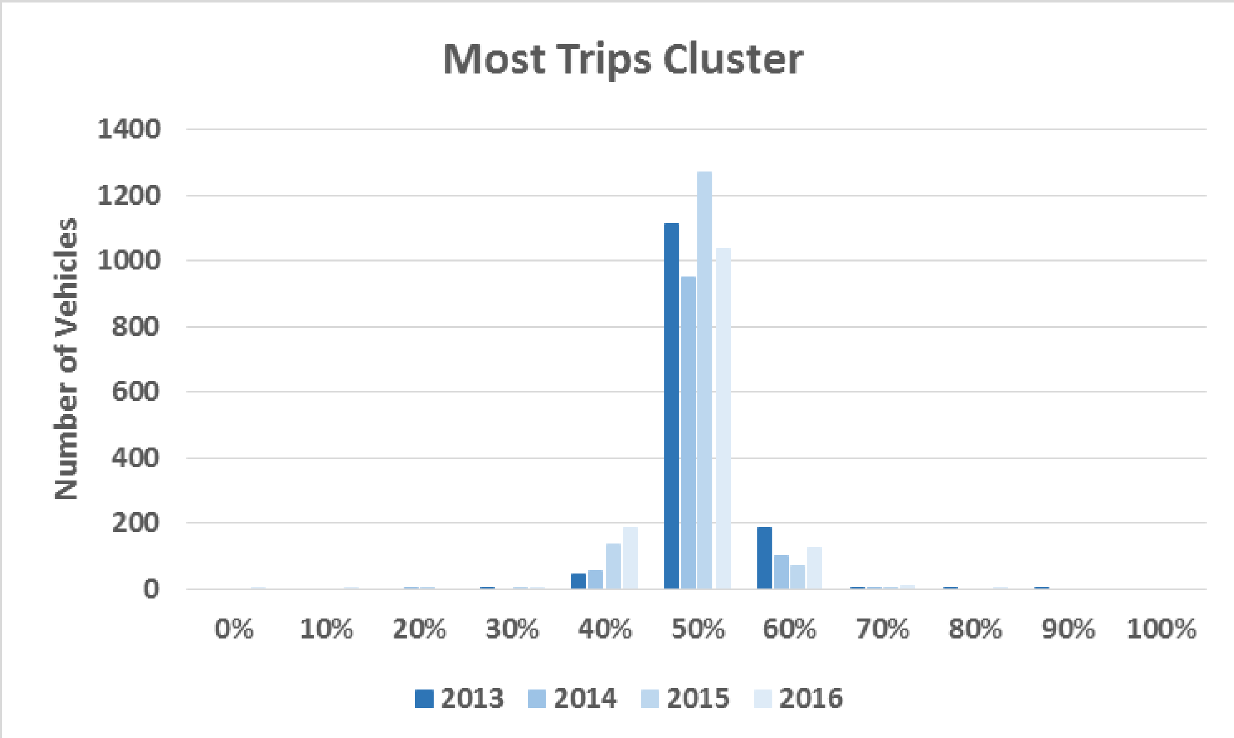
Temporal Efficiency Results

This section presents temporal efficiency measures for the Chicago taxi fleet. Figure 4-6 presents histograms of temporal efficiency for the taxis in the *most trips* and *normal trips* clusters. The x-axis represents the percentage of loaded time in bin intervals of 10%, and the y-axis is the number of taxis in each bin.

Figure 4-6 shows that around 50% of taxi in-service time is unproductive; i.e. not spent transporting users and collecting fares. For the taxis in the *most trips* cluster, the most productive taxis are productive 70-80% of their in-service time. The least productive taxis are productive 20-40% of their in-service time. However, a large majority of the taxis are productive 40-60% of their in-service time.

Like the taxis in the *most trips* cluster, only a few of the taxis in the *normal trips* cluster are productive 70-80% of their in-service time. A large majority of the *normal trips* cluster taxis were productive 40-60% of their in-service time. Figure 4-6 also shows that in 2013 and 2015 many *normal trip* taxis were only productive 10-20% of their in-service time.

The results in Figure 4-6 indicate that most taxis in the Chicago fleet are highly inefficient. Most drivers spend nearly half of their in-service time not generating revenue. This suggests that there is significant room for improvement in terms of operational efficiency.



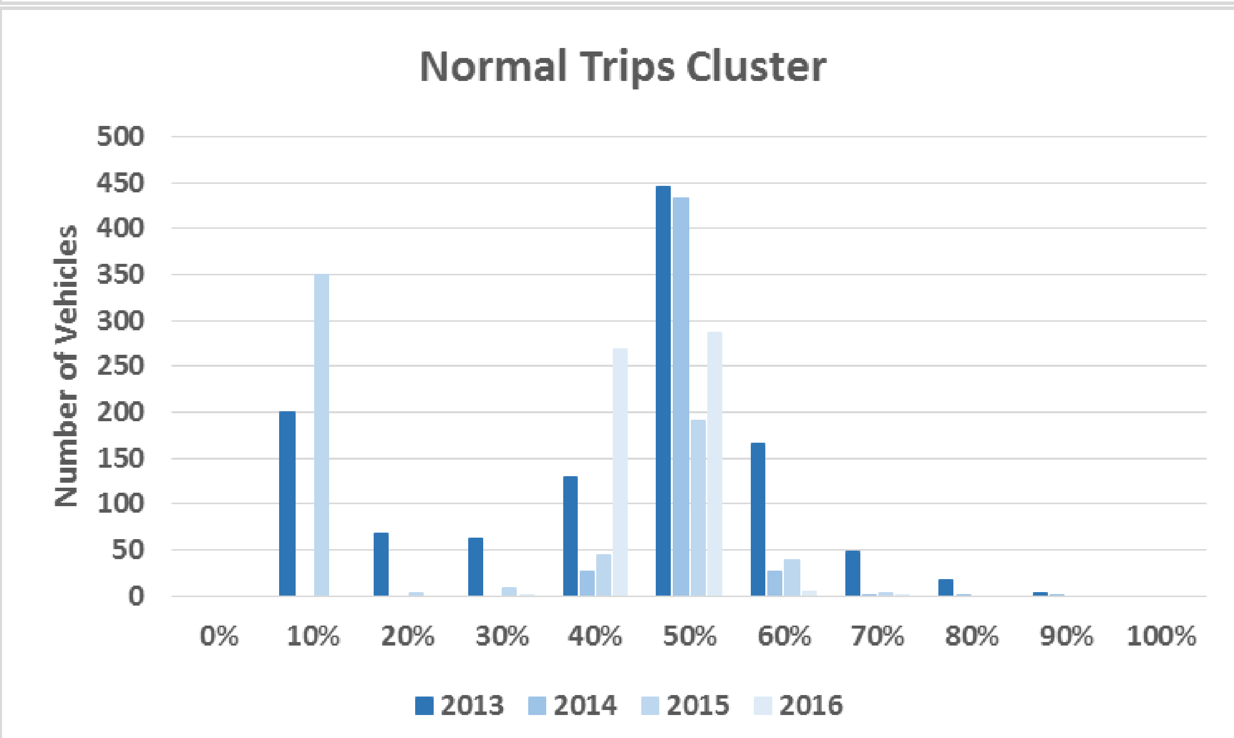


Figure 4-6: Histograms of the average temporal efficiency of taxis in the most (top) and normal (bottom) trips clusters by year

Spatial Efficiency Results

As an illustrative example, this section presents spatial efficiency measures for two taxis. Figure 4-7 displays a histogram of the spatial efficiency of every trip made by Taxi #44 between 2013 and 2016. Taxi #44 is a member of the *most trips* taxi cluster. The figure shows a clear gaussian distribution centered around 50% for each year. This indicates that around 50% of Taxi #44's miles are unproductive (i.e. empty).

Figure 4-7 also displays a histogram of the spatial efficiency of every trip made by Taxi #981between 2013 and 2016. Taxi #981, a member of the *less trips* taxi cluster, had a trip distribution slightly more uniform than Taxi #44; nevertheless, Taxi #981, on average, also seems to drive as many unproductive miles as productive miles. According to Figure 4-7, many of Taxi #981's trips have a spatial efficiency value between 10-30%; this is quite inefficient.

Like the temporal efficiency measures for the Chicago taxi fleet, the spatial efficiency measures for these two representative taxis indicate that there is significant room for improvement in terms of the operational efficiency of individual taxis. It is important to note that the histograms in Figure 4-7 exclude trips wherein $pos_{j,n,d} == pos_{j,n+1,p}$. This biases the overall results, because these trips would be highly efficient. However, the results are biased in the opposite direction because the empty distance ($e_{j,n}$) for each trip does not factor in any of the miles spent roaming for users.

Despite these issues with the data, Figure 4-7 show that many taxi trips are highly inefficient. As mentioned previously, this inefficiency negatively impacts the profit of taxi drivers and fleet operators, customer service quality, roadway congestion, and vehicle emissions.

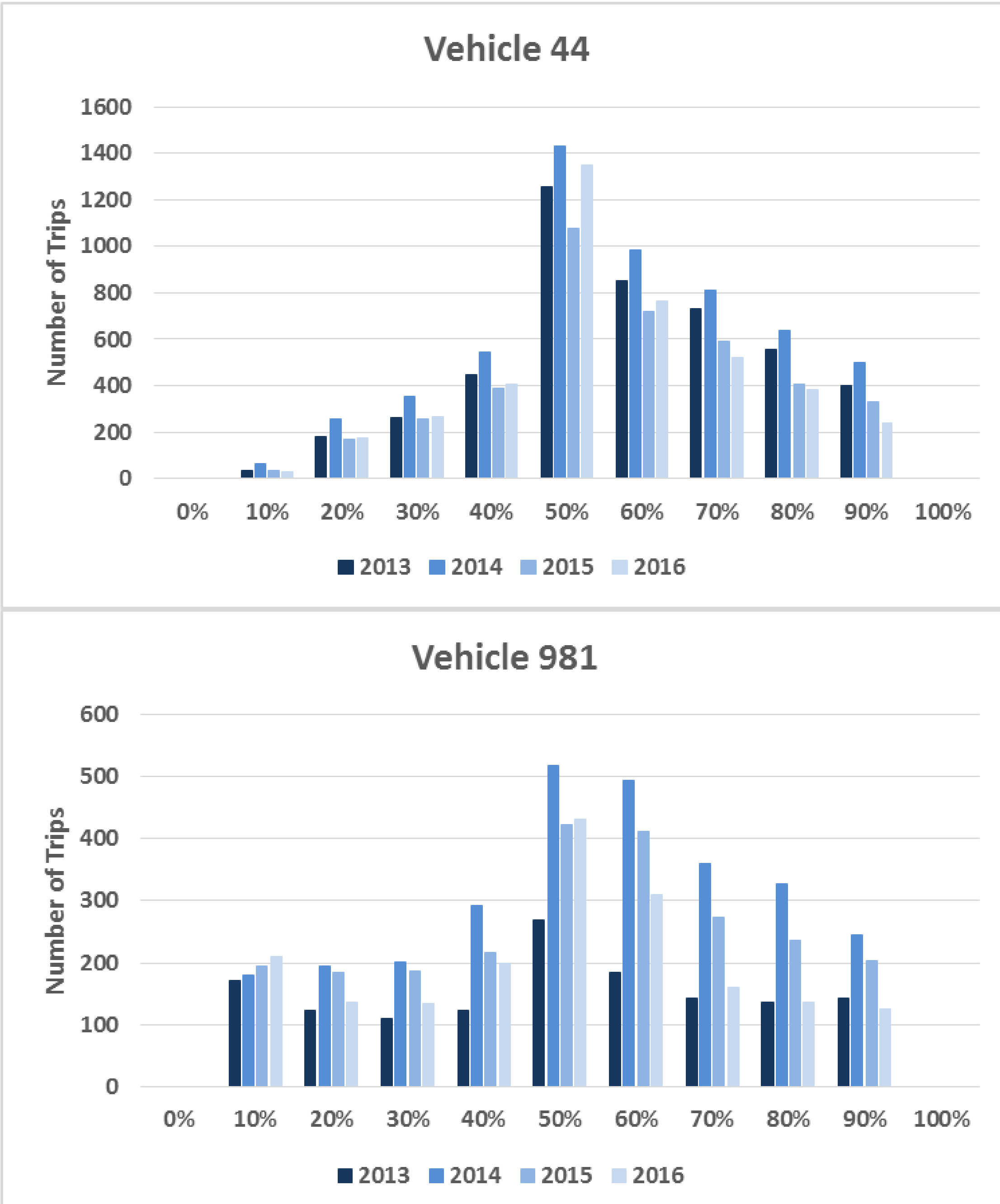


Figure 4-7: Histograms of the spatial efficiency of taxi #44 (top) and taxi #981's (bottom) trips from 2013 to 2016

4.5 Conclusion

This chapter used taxi data released by the city of Chicago to analyze the operational efficiency of taxis in the Chicago taxi fleet. The chapter presents two metrics to characterize and quantify the operational efficiency of individual taxis. The temporal efficiency metric determines the percentage of in-service time taxis spend transporting users. The spatial efficiency metric quantifies the percentage of a taxi's miles that are loaded. These metrics provide an effective means to measure the efficiency of individual taxis and taxi trips. Results indicate that most Chicago taxis and taxi-trips generate as many empty miles as productive miles; i.e. only 40-60% of taxi fleet miles are productive. Similarly, the temporal efficiency results indicate that taxis are unproductive 40-60% of their in-service time.

The efficiency results indicate that there is significant room for improvement in the efficiency of the Chicago taxi fleet. If SAMS providers can operate their fleets more efficiently than Chicago taxis, they can pass these cost savings onto users and provide lower cost transportation to users. From a societal perspective, the spatial inefficiency of taxis is likely increasing traffic congestion and generating extra vehicle emissions.

Future research directions include further development of algorithms to filter taxi trips, estimating missing data values, and better handling of the spatial and temporal trip aggregation, to better evaluate the spatial and temporal efficiency of taxis.

Chapter 5 Modeling Framework

5.1 Overview

This chapter presents an overview of the framework employed in this thesis to model on-demand SAMS operational problem. As mentioned previously, on-demand SAMS operational problems are stochastic dynamic vehicle routing problems. In fact, the on-demand SAMS operational problems in this thesis are highly-dynamic, stochastic, and quite large. The modeling framework in this chapter was developed to capture the state of the system, its evolution over time (through transition functions), sequential decision-making, and introduction of exogenous information. The framework and the mathematical notation are similar to Markov decision process (MDP) models (Puterman, 2014); however, the models in this thesis do not have all the properties of MDPs.

5.2 Model Components

The modeling framework contains six elements, namely, decision epochs, state variables, decision variables, exogenous information, transition function, and objective function. This section provides a description of these six elements in the broad context of on-demand SAMS operational problems.

Decision Epochs

Decision epochs represent the points in time when the SAMS fleet controller makes decisions. Decision epochs can be determined exogenously, or they can be a function of the system state. This thesis assumes a finite horizon with a pre-defined number of decision epochs. Let K be the set of decision epochs, |K| the number of decision epochs, and k the index of the kth decision epoch. The time between decision epochs is a fixed value, denoted I^d and referred to as the inter-decision time. Hence, decisions are made at regular intervals rather than as requests enter the system or AVs complete trips. The regular intervals serve two purposes. First, they allow user requests to queue before assigning them to AVs, which is often beneficial. Second, as solving the decision problem can take a few seconds, it is necessary to make sure the inter-decision time interval is greater than the solution time. The variable $t_k \in T$ denotes the time of decision epoch $k \in K$.

State Variables

The state variable S_k contains all the information necessary to model the system from the current epoch $k \in K$, to the end of the modeling period $|K| \in K$. State variables for SDVRPs are often multi-dimensional because it is necessary to keep track of the location and status of AVs and users. It is also sometimes necessary to keep track of the state of subregions; e.g. how many empty AVs are in each subregion.

Decision Variables

At each decision epoch k, given the state of the system S_k, the SAMS fleet controller can control the system via changing the plans of AVs. Let X_k denote the set of decision variables at decision epoch k. In on-demand SAMS operational problems, the decision variable X_k can also be multi-dimensional. Typical decisions include the assignment of AVs to open user requests, and the assignment of empty AVs to subregions.

It is also necessary to identify constraints for decision variables. For example, AVs should not be assigned to pick up more than one user request at each decision epoch k and user requests should not be assigned to more than one AV.

Exogenous Information

Between each decision epoch k, the stochastic dynamic system changes, as exogenous information enters the system and exogenous events change the state of the system. In the on-demand SAMS operational problems in this thesis, the main form of exogenous information are the user requests. A second form of exogenous information in one of the SAMS operational problems in this thesis are the set of times users actually release their carsharing vehicles back to the fleet controller. In MDP models, the exogenous information that enters the system between decision epochs $k-1$ and k is typically denoted ω_k. The models in this thesis will use the same notation.

Transition Function

The transition function defines how the state of the system S_k updates from decision epoch k to the next decision epoch $k+1$. For example, the decision epoch time t_k updates as follows, $t_{k+1} = t_k + I^d$. The decision variables and exogenous information impact the transition functions. For example, if one of the state variables is the status of a user and a decision is made to assign an AV to this user requests at epoch k, then at epoch $k+1$, the status of the user must be updated to indicate she has been assigned. This process must be captured through a transition function. Identifying relevant transition functions is especially helpful for coding discrete-event computer simulations of SAMS operational problems.

Objective Function

Let $C(S_k, X_k)$ denote the cost of being in state S_k and making decision X_k. For control problems such as SDVRPs, the solution is a policy $\pi \in \Pi$. Each policy π maps states to decisions; i.e. given S_k, policy $\pi \in \Pi$ yields decision $X_k^\pi(S_k)$. The objective of a SDVRP is to determine an optimal policy $\pi^* \in \Pi$ that minimizes the objective function in Eqn. (7), subject to the constraints on the decision variables.

$$\min_{\pi \in \Pi} E^\pi \left[\sum_{k \in \mathrm{K}} C\big(S_k, X_k^\pi(S_k)\big) \right] \tag{7}$$

Unfortunately, the very large (i.e. high-dimension) state space for the on-demand SAMS operational problems in this thesis make Eqn. (7), analytically intractable due to the curse of dimensionality (Powell, 2011). Researchers typically approximate the problem to obtain solutions.

5.3 Conclusion

The modeling framework presented in this chapter is employed throughout the thesis to model and optimize SAMS operational problems as well as simulate and analyze the efficiency of SAMS fleets. The six components of the modeling framework include, decision epochs, state variables, decision variables, exogenous information, transition functions, and the objective function. These model components effectively capture the dynamic and stochastic nature of the SAMS fleet operational problems this thesis addresses.

Chapter 6 On-demand SAMS without Shared Rides[6]

6.1 Overview

This chapter presents the *on-demand SAMS without shared rides* and its operational problem. The service is defined by a fleet of AVs, controlled by a central operator that provides direct origin-to-destination service to users who request rides via a mobile application and expect to be picked up within a few minutes. This service is similar to existing ridesourcing services offered by Uber and Lyft, except (i) the vehicles are driverless, (ii) fleet size is fixed, and (iii) the SAMS provider has complete control over the AVs.

The operational problem associated with the *on-demand SAMS without shared rides* is a stochastic dynamic control problem. This chapter also assumes that the SAMS fleet controller has no deterministic or stochastic information about future user requests. The SAMS controller must assign AVs to open user requests in real-time as user requests enter the system dynamically and randomly (i.e. without the fleet controller's prior knowledge). As there is likely no optimal policy for this sequential stochastic control problem, this chapter presents and compares six AV-user assignment strategies (i.e. control policies). The results show that optimization-based AV-user assignment strategies, strategies that allow en-route pickup AVs to be diverted to new user requests, and strategies that incorporate en-route drop-off AVs in the assignment problem, reduce

[6] This chapter parallels (Hyland and Mahmassani, 2018)

fleet miles and decrease user wait times. The more-sophisticated AV-user assignment strategies significantly improve operational efficiency when fleet utilization is high (e.g. during the morning or evening peak); conversely, when fleet utilization is low, simply assigning user requests sequentially to the nearest idle AV is comparable to more-advanced strategies. Simulation results also indicate that the spatial distribution of user requests significantly impacts the empty fleet miles generated by the on-demand SAMS.

6.2 Motivation and SAMS Definition

This chapter focuses on the real-time operation of a SAMS fleet. The most relevant operational-level advantage of AVs is their ability to safely and near-instantaneously receive and execute changes in vehicle plans (e.g. routes, schedules, and traveler assignments) coming from the fleet controller. From a fleet management perspective, the principal advantage of AVs is their guaranteed compliance with these real-time plan changes, and more generally the fleet manager's operational policies.

Motivated by the cost and performance benefits of AVs described in Chapter 1, the ability of SAMS fleet operators to completely control individual vehicles, and the importance of operational efficiency in terms of the success of on-demand SAMSs, this chapter addresses the underlying operational problem associated with an *on-demand SAMS without shared rides.* The SAMS's characteristics are as follows:

- Users request a pickup location and a drop-off location via a smartphone application;
- Users want AVs to come to their pickup location immediately;

- Users will always be served, if they are willing to wait, i.e. the fleet controller cannot reject user requests;
- AVs transport users directly from their requested pickup location to their drop-off location, i.e. no en-route detours to pick up or drop off other users;
- AVs in the fleet are functionally homogeneous;
- The fleet size is fixed in the short term (i.e. one-day);
- The fleet controller has complete control over each AV.

The fleet size is assumed to be fixed as ridesourcing companies, technology companies, and car manufacturers have stated that they plan to provide mobility services with their own AV fleet, rather than sell individual AVs to users (Waymo, 2017; Wingfield, 2017).

The *on-demand SAMS without shared rides* operational problem is a stochastic dynamic control problem (SDCP). The problem is dynamic because decisions need to be made sequentially as new user requests enter the system. The problem is stochastic because future user requests are unknown. The goal of the SAMS fleet controller is to serve the user requests as efficiently as possible. The specific objectives are to minimize user wait times and SAMS fleet miles. Short user wait times are likely to be a key factor in the success of AMOD services (Krueger et al., 2016), especially if users are to forego owning their own vehicles and rely on AMOD services for all travel.

This study assumes the fleet controller has no stochastic information about future requests in order to focus on the efficiency of dispatching/assignment strategies. When stochastic information is available, the operational problem involves AV-user assignment and AV repositioning. In this case, it is difficult to dissociate the AV-user assignment problem from the larger operational problem that involves both AV-user assignment and AV repositioning.

6.3 Problem Statement and Model

This section presents a formal description of the *on-demand SAMS without shared rides* operational problem and then presents a mathematical model of the stochastic dynamic control problem.

Problem Statement

The *on-demand SAMS without shared rides* operational problem is characterized by a fleet of AVs $\mathcal{V} = \{V_1, V_2, \dots V_j, V_{j+1}, \dots, V_{|\mathcal{V}|}\}$ that aim to serve users $\mathcal{C} = \{C_1, C_2, \dots C_i, C_{i+1} \dots, C_{|\mathcal{C}|}\}$ who request service during the finite time horizon $T = [0, t_{max}]$, over a rectangular geographic service region $\mathcal{G}$ with side lengths L_1 and L_2. The geographical region $\mathcal{G}$ is a Manhattan plane $\mathcal{G} = \{(x, y) |\, x \in [0, L_1], y \in [0, L_2]\}$. The distance between any two locations l_1 and l_2, where $l_1, l_2 \in \mathcal{G}$, is denoted $d(l_1, l_2)$.

At time $t = 0$, AVs may be located at one or several depots, or they may be dispersed throughout the entire region. User requests occur according to an unknown stochastic process $\mathcal{F}^C$. Each user request C_i comes with a request time $t_r^{C_i} \in T$, pickup location $l_p^{C_i} \in \mathcal{G}$, and drop-off location $l_d^{C_i} \in \mathcal{G}$. The AV must pick up the user at her requested pickup location. Let $t_p^{C_i}$ and $t_d^{C_i}$ denote the time an AV picks up and drops off user C_i, respectively.

The goal of the SAMS fleet controller is to efficiently serve the user requests via minimizing user wait times and empty fleet distance. Let $d_e^{V_j}$ and d^{V_j} denote the empty distance and total distance of AV V_j, respectively. The operational efficiency of the fleet will be determined based on the percentage of empty fleet miles, $\frac{\sum_{V_j \in \mathcal{V}} d_e^{V_j}}{\sum_{V_j \in \mathcal{V}} d^{V_j}}$, and average user wait $\frac{\sum_{C_i \in \mathcal{C}} \left(t_p^{C_i} - t_r^{C_i}\right)}{|\mathcal{C}|}$.

Model

This chapter makes the following modeling assumptions:

- The fleet size is fixed in the short term (i.e. one-day)
- Users will wait indefinitely to be served

Given these assumptions, and the problem statement presented above, this section presents a model of the SAMS operational problem. The model includes six elements, namely, decision epochs, states, decisions, exogenous information, transition function, and objective function.

Decision Epochs

This chapter assumes a finite horizon with a pre-defined number of decision epochs. Let K be the set of decision epochs, $|\mathrm{K}|$ the number of decision epochs, and k the index of the k$^{\text{th}}$ decision epoch. The time between decision epochs is a fixed value, denoted I^d and referred to as the inter-decision time. The variable $t_k \in T$ denotes the time of decision epoch $k \in K$.

State Variables

For the *on-demand SAMS without shared rides* operational problem, the system state at epoch k (S_k) includes several dimensions. This chapter delineates two sets of entities – users and AVs – with states that need to be updated. Let the states of users and AVs at epoch $k \in K$ be denoted $S_k^{\mathcal{C}}$, and $S_k^{\mathcal{V}}$ respectively, where S_k is completely defined by the set $\left(S_k^{\mathcal{C}}, S_k^{\mathcal{V}}, \mathrm{t_k}\right)$.

User State. The user state $S_k^{\mathcal{C}}$ is the tuple $\left(\sigma_k^{\mathcal{C}}, w_k^{\mathcal{C}}, \mathcal{C}\right)$ denoting the status, elapsed wait time, and static information of users, respectively. The static user information includes user request times $t_r^{\mathcal{C}}$, user pickup locations $l_p^{\mathcal{C}}$, and user drop-off locations $l_d^{\mathcal{C}}$.

For each user C_i, user status $\sigma_k^{C_i}$ takes on a value in the set {0,1,2,3}:

$$\sigma_k^{C_i} = \begin{cases} 0, & C_i \text{ has not requested service by time } t_k \\ 1, & C_i \text{ has requested service but has not been assigned by } t_k \\ 2, & C_i \text{ has been assigned but not picked up by } t_k \\ 3, & C_i \text{ has been picked up} \end{cases}$$

If user C_i has not requested service at t_k (i.e. $\sigma_k^{C_i} = 0$), then the static information associated with user C_i is unknown to the fleet controller. The elapsed wait time of a user $w_k^{C_i}$ is simply the difference between the current time and the user's request time: $w_k^{C_i} = t_k - t_r^{C_i}$.

AV State. The AV state $S_k^{\mathcal{V}}$ is the tuple $\left(\varphi_k^{\mathcal{V}}, l_k^{\mathcal{V}}, r_k^{\mathcal{V}}\right)$ denoting the status, location, and route plan of every AV V_j at decision epoch k, respectively. For each AV V_j, $\varphi_k^{V_j}$ takes on a value in the set {0,1,2,3}:

$$\varphi_k^{V_j} = \begin{cases} 0, & V_j \text{ is idle at time } t_k \\ 1, & V_j \text{ is enroute to pick up a user at } t_k \\ 2, & V_j \text{ is enroute to drop off a user at } t_k \end{cases}$$

The operational strategies presented in this chapter vary the AVs included in the assignment problem, at decision epoch k, based on their statuses $\varphi_k^{V_j}$.

The set of AV routes $r_k^{\mathcal{V}} = \left(r_k^{V_1}, r_k^{V_2}, \dots, r_k^{V_j}, \dots, r_k^{V_{|\mathcal{V}|}}\right)$ provides the sequenced set of locations AVs will visit next, at decision epoch k. In this chapter, at every epoch k, an AV route $r_k^{V_j}$ can only include a maximum of two locations because the model and solution approach only allow for an AV to be assigned to serve one unserved user.

Decisions

At each decision epoch k, given the state of the system S_k, the SAMS fleet operator can control the system via changing the plans of AVs. Let X_k denote the set of decision variables at decision epoch k. To model the decision problem, this chapter introduces variables x_k^{ij}, defined as follows:

$$x_k^{ij} = \begin{cases} 1, & \text{if AV } V_j \text{ is assigned to pick up user } C_i \text{ at time } t_k \\ 0, & \text{otherwise} \end{cases}$$

There are four constraints on the decision variables, displayed in Eqn. (8)-(11).

$$\sum_i x_k^{ij} \leq 1 \qquad \forall j, k \tag{8}$$

$$\sum_j x_k^{ij} \leq 1 \qquad \forall i, k \tag{9}$$

$$x_k^{ij} = 0 \qquad \forall i \notin \mathcal{C}', \forall j \notin \mathcal{V}' \tag{10}$$

$$x_k^{ij} \in \{0,1\} \qquad \forall i, j, k \tag{11}$$

The constraint in Eqn. (8) ensures each AV V^j is assigned to at most one open user request C^i. The constraint in Eqn. (9) ensures that no more than one AV is assigned to an open user request C^i. The constraint in Eqn. (10) only allows available users, denoted $\mathcal{C}'$, and available AVs, denoted

$\mathcal{V}'$ to be assigned. The solution approaches presented in the next section vary the users and AVs that are considered 'available' in the problem based on their statuses. The constraint in Eqn. (11) ensures the decision variables take on binary values.

Exogenous Information

The problem includes one source of exogenous information, namely, the user requests. Hence, let $\omega_{k+1} = \gamma_k$, where γ_k is the set of previously unrequested user requests ($\sigma_k^{\mathcal{C}} = 0$) with a request time between t_k and t_{k+1}; i.e. $\gamma_k \subseteq \{C_i | \sigma_k^{C_i} = 0, t_k < t_r^{C_i} \leq t_{k+1}\}$.

Transition Function

The transition function defines how the state of the system S_k updates from decision epoch k to the next decision epoch $k+1$. The decision epoch time t_k updates as follows:

$$t_{k+1} = t_k + I^d$$

The user state $S_k^{\mathcal{C}} = (\sigma_k^{\mathcal{C}}, w_k^{\mathcal{C}}, \mathcal{C})$ contains two elements that need to be updated The user information ($\mathcal{C}$) stays the same; whereas the users' statuses ($\sigma_k^{\mathcal{C}}$) and elapsed wait times ($w_k^{\mathcal{C}}$) need to be updated. Elapsed wait time updates as follows, $w_{k+1}^{C_i} = w_k^{C_i} + I^d$. Updating user statuses is more complex and depends on the AVs and open user requests considered in the AV-user assignment strategy. Let $\mathbb{1}_{C_i \in \gamma_k}$ denote an indicator variable equal to 1 if $C_i \in \gamma_k$ (i.e. user C_i has a request time in the current epoch). Things that need to be considered in the transition function are whether the user requests service between decision epochs ($\mathbb{1}_{C_i \in \gamma_k}$), whether users are assigned or reassigned to AVs (x_k^{ij}), whether an AVs picks up a user ($\mathbb{1}_{t_k \leq t_p^{C_i} < t_{k+1}}$), or drops off a user ($\mathbb{1}_{t_k \leq t_d^{C_i} < t_{k+1}}$) in the current epoch.

The AV state $S_k^{\mathcal{V}} = \left(\varphi_k^{\mathcal{V}}, l_k^{\mathcal{V}}, r_k^{\mathcal{V}}\right)$ contains three elements that all need to be updated. The location of AV V_j updates as follows, $l_{k+1}^{V_j} = l_k^{V_j} + vI^d \Lambda_{l_k^{\mathcal{V}}}^{r_k^{V_j}(1)}$, where v is the vehicle speed, I^d is the length of the epoch, and $\Lambda_{l_k^{V_j}}^{r_k^{V_j}(1)}$ is the unit direction between AV V_j's current location ($l_k^{V_j}$) and the next stop on its route, $r_k^{V_j}(1)$. If the vehicle does not have a next stop, the vehicle remains in its current location

In the case where reassignments are not allowed, AV route plans are updated as follows, $r_{k+1}^{V_j} = r_k^{V_j} \cup \left\{l_p^{C_i}, \quad \text{if } x_k^{ji} = 1, l_p^{C_i} \notin r_k^{V_j}\right\}$. In the case where reassignments are allowed, stops may need to be removed from $r_k^{V_j}$. The route plans are also updated when a vehicle reaches the next stop in its route, $r_{k+1}^{V_j} = r_k^{V_j} \backslash \left\{l_p^{C_i}, \quad \text{if } l_k^{V_j} = l_p^{C_i}\right\}$.

AV status ($\varphi_k^{\mathcal{V}}$) transitions also depend on the AV-user assignment strategy, the decision variable (x_k^{ji}) and location of the AV and the user's pickup location, $\mathbb{1}_{\left(l_k^{V_j} = l_p^{C_i}\right)}$, as well as the user's drop-off location $\mathbb{1}_{\left(l_k^{V_j} = l_d^{C_i}\right)}$.

Objective Function

Let $C(S_k, X_k)$ denote the cost of being in state S_k and making decision X_k. For SDCPs, the is a policy $\pi \in \Pi$. Each policy π maps states to decisions; i.e. given S_k, policy $\pi \in \Pi$ yields decision $X_k^{\pi}(S_k)$. The objective of an SDCP is to determine an optimal policy $\pi^* \in \Pi$ that minimizes the objective function in Eqn. (12), subject to the constraints in Eqn. (8)-(11) on the decision variables.

$$\min_{\pi \in \Pi} E^{\pi}\left[\sum_{k \in \mathrm{K}} C\left(S_k, X_k^{\pi}(S_k)\right)\right] \tag{12}$$

Unfortunately, the very large (i.e. highly-dimensioned) state space for the *on-demand SAMS without shared rides* operational problem makes Eqn. (12) analytically intractable due to the curse of dimensionality (Powell, 2011). Researchers typically approximate the problem to obtain solutions. Moreover, as the fleet controller does not have stochastic information about future requests, the impact of decision made at epoch k on future epochs is unknown. Hence, to solve the *on-demand SAMS without shared rides* operational problem with no stochastic information, the fleet operator solves the local problem $\min_{x_k}\{C(S_k, \mathrm{x_k})\}$ at every epoch. This is typically referred to as a myopic policy.

6.4 Solution Approach

This section details the solution approach to solve the *on-demand SAMS without shared rides* operational problem where the fleet controller has no stochastic information. The next subsection formulates the local cost problem $\min_{x_k}\{C(S_k, \mathrm{x_k})\}$ and the following section presents polices for the full SDCP $\min_{\pi \in \Pi} E^{\pi}\left[\sum_{k \in \mathrm{K}} C\left(S_k, X_k^{\pi}(S_k)\right)\right]$ via changing AVs and users considered in the local cost problem.

AV-User Assignment Problem

This subsection presents a mathematical formulation of the local cost problem $\min_{x_k}\{C(S_k, \mathrm{x_k})\}$, referred to as the AV-user assignment problem. Let d_k^{ij} denote the distance between the C_i's pickup location ($l_p^{C_i}$) and V_j's current location ($l_k^{V_j}$). From the decision variable section above, $\mathcal{C}'$ and $\mathcal{V}'$ denote the sets of users and AVs that are available for assignment, respectively. These sets of 'available' users and AVs vary as a function of the assignment strategy. Since, the problem is AV-user assignment problem is local, the fleet controller only solves for the current decision epoch k; for notational simplicity, the index k is removed from the formulation of the mathematical program.

The AV-user assignment formulation depends on whether, in the dynamic system at time k, the number of users included in the assignment ($|\mathcal{C}'|$) is less than or greater than the number of AVs in the assignment ($|\mathcal{V}'|$). The mathematical programming formulation for the AV-user assignment problem when the number of available users is greater than the number of available AVs ($|\mathcal{C}'| > |\mathcal{V}'|$) is given in Eqn. (13)-(16).

$$\min_{x^{ij}} \sum_{i \in \mathcal{C}'} \sum_{j \in \mathcal{V}'} \left(d^{ij} x^{ij} - \gamma w^i x^{ij} \right) \tag{13}$$

s.t.

$$\sum_{j \in \mathcal{V}'} x^{ij} \leq 1 \qquad \forall i \in \mathcal{C}' \tag{14}$$

$$\sum_{i \in \mathcal{C}'} x^{ij} = 1 \qquad \forall j \in \mathcal{V}' \tag{15}$$

$$x^{ij} \geq 0 \qquad \forall i \in \mathcal{C}', \forall j \in \mathcal{V}' \tag{16}$$

Equation (14) ensures that each user is assigned to at most one AV. Equation (15) ensures each AV is assigned to a user. Equation (16) requires the decision variable, x_{ij}, to be non-negative. However, because the constraint matrix is totally unimodular, x^{ij} will only take on integer values. The objective function in Eqn. (13) has two terms. The first term represents the total distance between users and the AVs that they are assigned to. The second term represents the elapsed wait time of the assigned users. The parameter γ weights the relative importance of assigning AVs to users that have been waiting a long time; γ also converts the time units associated with w^i to the distance units associated with d^{ij}. Given that all users are not assigned to an AV because $|\mathcal{C}'| > |\mathcal{V}'|$, the second term incentivizes the SAMS fleet controller to assign AVs to users with large elapsed wait times (w^i). Without this second term, the fleet controller will not assign AVs to users in the periphery of the service region, if they are all busy serving user requests in the core of the service region. The challenge of serving user requests in the periphery is quite common and was recognized in one of the seminal DRP studies (Psaraftis, 1980). The objective function in Eqn. (13) handles this challenge elegantly and effectively. Although this formulation does not guarantee the periphery users will be assigned at the current epoch k, the longer the users in the periphery wait, the higher the incentive to pick up these travelers from the perspective of the fleet operator.

The mathematical programming formulation for the AV-user assignment problem changes slightly when the number of available users is less than the number of available AVs ($|\mathcal{C}'| \leq |\mathcal{V}'|$). This problem is formulated in Eqn. (17)-(20).

$$\min_{x^{ij}} \sum_{i \in \mathcal{C}'} \sum_{j \in \mathcal{V}'} \left(d^{ij} x^{ij} \right) \tag{17}$$

s.t.

$$\sum_{j \in \mathcal{V}'} x^{ij} = 1 \qquad \forall i \in \mathcal{C}' \tag{18}$$

$$\sum_{i \in \mathcal{C}'} x^{ij} \leq 1 \qquad \forall j \in \mathcal{V}' \tag{19}$$

$$x^{ij} \geq 0 \qquad \forall i \in \mathcal{C}', \forall j \in \mathcal{V}' \tag{20}$$

Because the number of users is less than the number of AVs, and the constraint in Eqn. (18) requires each user to be assigned to an AV, the second term in the objective function of Eqn. (13) is no longer relevant. Every user, including those in the periphery, will be assigned to an AV due to the constraint in Eqn. (18). The objective in Eqn. (17) is to minimize the overall distance between users and the AVs they are assigned. Equation (19) ensures that each AV is assigned to at most one user. Finally, Eqn. (20) requires the decision variable, x^{ij}, to be non-negative.

The solution algorithm determines whether to solve the AV-user assignment problem in Eqn. (13)-(16) or Eqn. (17)-(20) depending on the number of number of available users ($|\mathcal{C}'|$) and the number of available AVs ($|\mathcal{V}'|$). These two math programming formulations provide a baseline model for assigning AVs to user requests. The next subsection presents six unique assignment strategies to solve the SDCP. Four of the strategies employ the AV-user assignment problem; the formulation varies slightly, but significantly across the four strategies. For example, the simpler strategies only include unassigned users $\mathcal{C}' = \left\{C_i | \sigma_k^{C_i} = 1\right\}$ and idle AVs $\mathcal{V}' = \left\{V_j | \varphi_k^{V_j} = 0\right\}$. Let V_k^I, V_k^P, V_k^D be the sets of idle AVs ($\varphi_k^{V_j} = 0$), pickup AVs ($\varphi_k^{V_j} = 1$), and drop-off AVs ($\varphi_k^{V_j} = 2$) at epoch k, respectively, and V^I, V^P, V^D for short. Similarly, let C^U, C^A, C^V be the sets of unassigned users ($\sigma_k^{C_i} = 1$), assigned users ($\sigma_k^{C_i} = 2$), and in-vehicle users ($\sigma_k^{C_i} = 3$) at epoch k, respectively.

AV-User Assignment Strategies

This section presents six different AV-user assignment strategies. To solve the *on-demand SAMS without shared rides* operational problem, the SAMS fleet controller repeatedly solves the local optimization problem (i.e. the AV-user assignment problem) based on the state of users and AVs at epoch k.

The first two strategies are simplistic first-come, first-served (FCFS) strategies, whereas strategies three through six employ the AV-user mathematical programming formulation. The difference between the four strategies comes down to the AVs ($\mathcal{V}'$) and users ($\mathcal{C}'$) that are treated as 'available' in the problem formulation.

Figure 6-1 through Figure 6-6 display a toy example. The left side of each figure displays the assignment of AVs to users at time t_k and the right side displays the updated assignment of AVs to users at time $t_k + I^d$. The solid lines indicate assignments made at time t_k, and the dashed lines represent assignments or reassignments made at $t_{k+1} = t_k + I^d$. The triangles represent user drop-off locations and the squares represent pickup locations, wherein the dashed-line squares represent new user requests that enter the system between time t_k and time $t_k + I^d$.

Strategy 1

The first strategy assigns users FCFS to the longest idle AV (AV *1* was idling the longest when user *1* made a request, AV 2 had the second longest idle time, AV 3 the third longest idle time, etc.). Figure 6-1 illustrates the inefficiency associated with this heuristic strategy. On the left side of Figure 6-1, despite being the farthest AV from user *2*, AV *2* was assigned to user *2* only because AV *2* was idle longer than the other AVs in the fleet. Between t_k and $t_k + I^d$, user *3* and user *4* enter the system sequentially and they are each assigned sequentially to the longest remaining idle AV.

The resultant assignment in Figure 6-1 confirms the inefficient logic associated with the first assignment strategy. In this toy problem, given that the bottom edge of Figure 6-1 through Figure

6-6 is 3.5 miles, the total fleet mileage required to drop off user *1* and pick up users *2-4* under Strategy 1 is 12.9 miles.

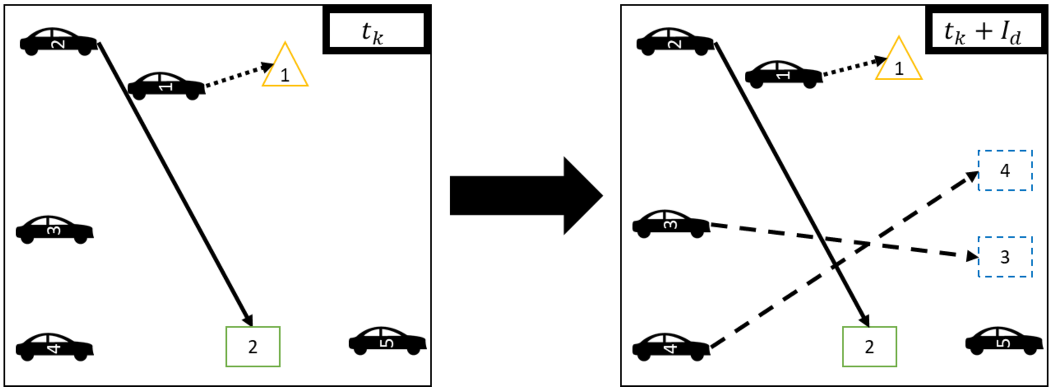


Figure 6-1: AV-user assignment based on Strategy 1 (total fleet miles = 12.9)

Strategy 2

The second strategy assigns users FCFS to the nearest idle AV. Assigning users FCFS is still an inefficient strategy; however, assigning them to the nearest idle AV should improve the fleet's operational efficiency, relative to Strategy 1. For example, on the left side of Figure 6-2, AV 5 is assigned to user 2, rather than the inefficient assignment of AV 2 to user 2 in Strategy 1.

On the right side of Figure 6-2, user 3 is assigned to AV 3, the nearest idle AV. Then, after user 3 is assigned, user 4 is assigned to AV 2, the nearest remaining idle AV. In Strategy 2, the users are assigned sequentially, rather than simultaneously between t_k and $t_k + I^d$. For the toy problem, the total fleet mileage associated with Strategy 2 is 9.7 miles, compared with 12.9 miles for Strategy 1. This is a significant improvement; however, assigning new sets of users simultaneously rather than FCFS or sequentially should further improve the fleet's efficiency.

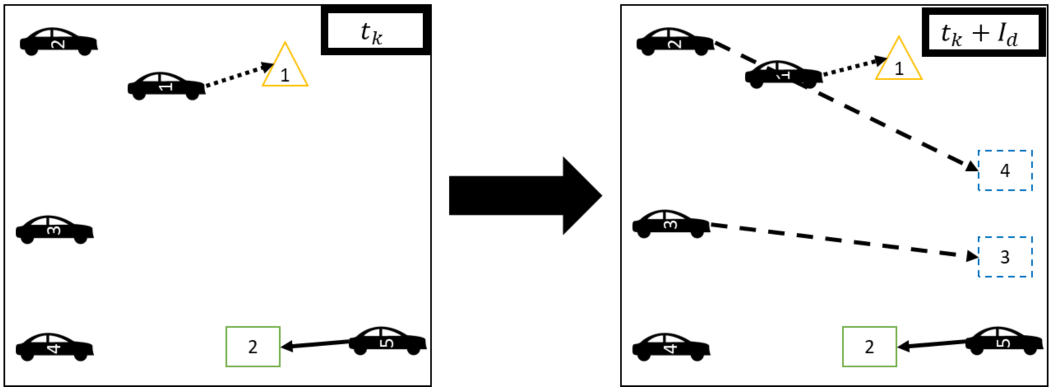


Figure 6-2: AV-user assignment based on Strategy 2 (total fleet miles = 9.7)

Strategy 3

In the third strategy, only unassigned users ($\mathcal{C}' = C^U$) and idle AVs ($\mathcal{V}' = V^I$) are considered in the AV-user assignment problem. In this strategy, once an AV-user assignment is made, it is not altered.

Figure 6-3 displays the results of the toy problem if Strategy 3 is employed. The assignment of AVs to users between t_k and $t_k + I^d$ is more efficient in Figure 6-3 than Figure 6-2 because user *3* and user *4* are assigned simultaneously, rather than sequentially. Simultaneous assignment allows the optimization solver to find the best assignment across both user 3 and user 4. The total fleet mileage associated with Strategy 3 in this toy problem is 9.0 miles.

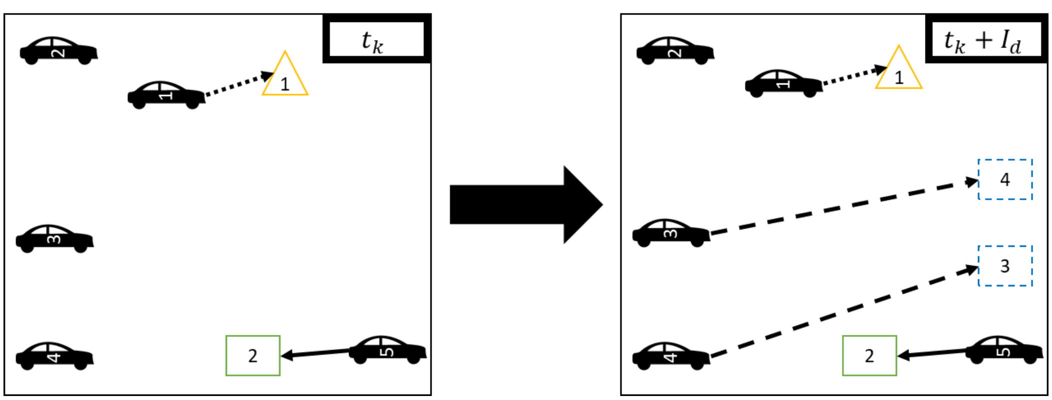


Figure 6-3: AV-user assignment based on Strategy 3 (total fleet miles = 9.0)

Strategy 4

In this strategy, unassigned and assigned users ($\mathcal{C}' = \{C^U \cup C^A\}$) as well as idle and en-route pickup AVs ($\mathcal{V}' = \{V^I \cup V^P\}$) are considered in the AV-user assignment problem. The inclusion of assigned users and en-route pickup AVs allows the reassignment of AVs to previously assigned users. In the dynamic fleet management literature, this reassignment of users is also known as vehicle diversion (Ichoua et al., 2006; Regan et al., 1995; Sheridan et al., 2013).

In the example problem in Figure 6-4, AV *5* is diverted from user *2* to user *3* between t_k and $t_k + I^d$. Additionally, user 2 is reassigned to AV *4* and AV *3* is assigned to user *4*. The total fleet mileage associated with Strategy 4 is 6.8 miles. Hence, for this problem instance, allowing user reassignment (or vehicle diversion) significantly reduces fleet miles.

Allowing reassignment is often a very beneficial strategy, as illustrated in Figure 6-4. However, without adding constraints to the AV-user math program, the strategy can result in unwanted outcomes, such as a previously assigned user no longer being assigned, and a user constantly being reassigned to different AVs.

To prevent previously assigned users from being completely unassigned, a constraint is added to the mathematical programming formulation of the AV-user assignment problem. Let a_i equal 1 if user $i \in C^A$, and zero otherwise. The following constraint prevents a user from going from assigned to unassigned:

$$\sum_{j \in V'} x^{ij} = 1 \quad \forall i \in C^A \qquad \text{or} \qquad \sum_{j \in V'} x^{ij} - a_i \geq 0 \quad \forall i \in \mathcal{C} \tag{21}$$

To prevent users from constantly being reassigned, a constraint is added that only allows one reassignment per user. Let m^{ij} equal 1, if pickup AV $j \in V^P$ is en-route to pick up assigned user $i \in C^A$. Let b_i equal 1 if assigned user $i \in C^A$ has previously been reassigned. The following constraint prevents a user from being reassigned more than once:

$$b_i\left(m^{ij} - x^{ij}\right) \leq 0 \quad \forall\, i \in C^A, \ \forall j \in V^P \tag{22}$$

In addition to these constraints, a penalty is added to the objective function for reassigning AVs. Let δ denote the penalty for assigning a user to an en-route pickup AV $j \in V^P$. Let q_j equal 1, if AV j is en-route to pick up a user ($j \in V^P$). The objective of the AV-user assignment problem in Eqn. (13) then becomes:

$$\min_{x^{ij}} \left(\sum_{i \in \mathcal{C}'} \sum_{j \in \mathcal{V}'} \left(d^{ij} x^{ij} - \gamma w^i x^{ij} + \delta\left(1 - m^{ij}\right) q_j x^{ij} \right) \right) \tag{23}$$

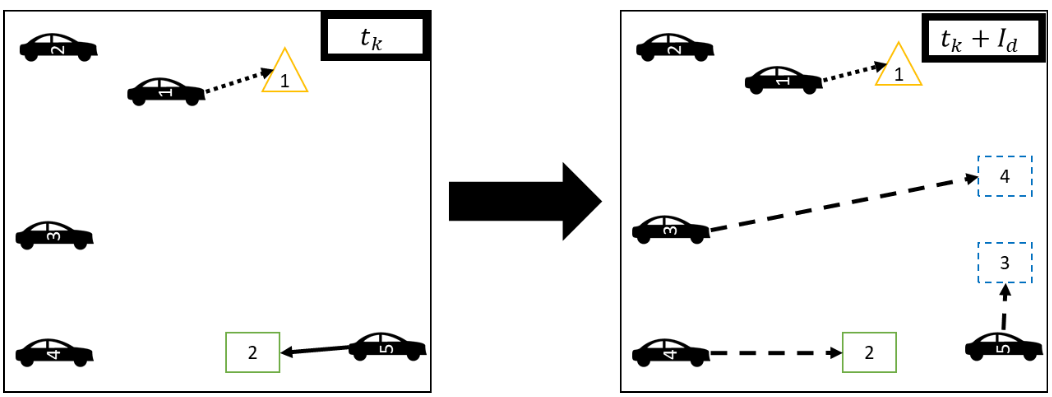


Figure 6-4: AV-user assignment based on Strategy 4 (total fleet miles = 6.8)

Strategy 5

In the fifth strategy, unassigned users ($\mathcal{C}' = C^U$) as well as idle and en-route drop-off AVs ($\mathcal{V}' = \{V^I \cup V^D\}$) are considered in the assignment problem; this is similar to the strategy employed in Maciejewski et al. (2016) for dispatching taxis. This strategy does not allow en-route pickup AVs to be diverted ($V^P \notin \mathcal{V}'$), nor does it allow users to be reassigned ($C^A \notin \mathcal{C}'$). However, this strategy essentially allows two-person schedules for AVs. That is, if a new user request $i' \in$

C^U enters the system, and an en-route drop-off AV $j' \in V^D$ can pick up user $i' \in C^U$ faster than all the other AVs, even after considering the remaining time/distance to drop off the user it is carrying $i'' \in C^V$, then the en-route drop-off AV $j' \in V^D$ is assigned to the new user request $i' \in C^U$.

The right side of Figure 6-5 is unique in that AV *1* is assigned to user *4* even though it was en-route to drop off user *1* at the time of assignment. For this problem instance, the total fleet mileage associated with Strategy 5 is 7.4 miles, an improvement over Strategy 3 (9.0 miles) but not quite as good as Strategy 4 (6.8 miles). It should be clear that combining Strategy 4 and Strategy 5, can further improve the fleet efficiency.

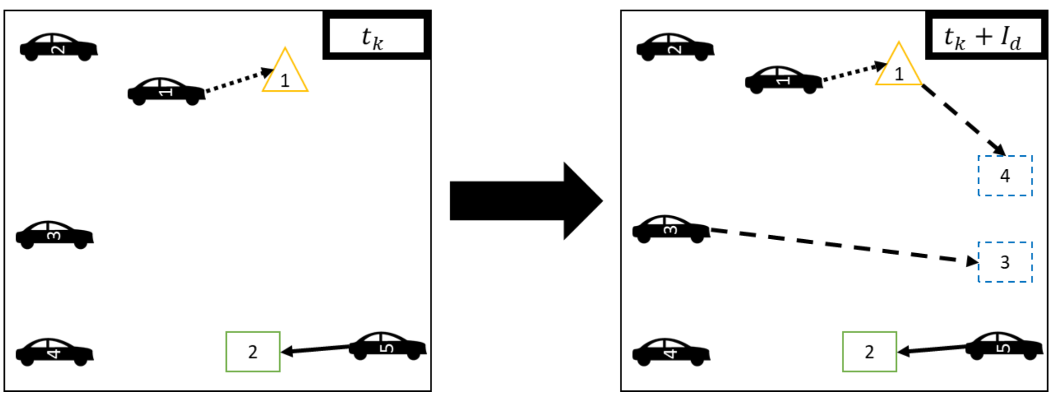


Figure 6-5: AV-user assignment based on Strategy 5 (total fleet miles = 7.4)

Determining d_{ij} for en-route drop-off AVs requires calculating two sets of distances. The distance between the current position of AV $j' \in V^D$ and the drop-off location of user $i'' \in C^V$: $dist\left(l_k^{j'}, l_d^{i''}\right)$ and the distance between the drop-off location of user $i'' \in C^V$ and the pickup location of user $i' \in C^A$: $dist\left(l_d^{i''}, l_p^{i'}\right)$. Therefore, $d_{ij} = dist\left(l_k^{j'}, l_d^{i''}\right) + dist\left(l_d^{i''}, l_p^{i'}\right)$.

As it takes time for a user to get out of an en-route drop-off AV, and the fleet controller probably only wants to assign a user to an en-route drop-off AV if it is a much better option than an idle AV, a penalty is added in the objective function for assigning an en-route drop-off AV to a user. Let p_j equal 1 if AV j is en-route to drop off a user ($j \in V^D$). Let φ denote the penalty of assigning a user to an en-route drop-off AV. The objective function in Eqn. (13), changes to:

$$\min_{x_{ij}} \left(\sum_{i \in \mathcal{C}'} \sum_{j \in \mathcal{V}'} \left(d^{ij} x^{ij} - \gamma w^i x^{ij} + \varphi p_j x^{ij} \right) \right) \tag{24}$$

Strategy 6

In the sixth and final strategy, unassigned and assigned users ($\mathcal{C}' = \{C^U \cup C^A\}$) as well as all AVs (idle, en-route pickup, and en-route drop-off, $\mathcal{V}' = \{V^I \cup V^P \cup V^D\}$) are considered in the assignment problem. Strategy 6 combines the valuable additions in Strategy 4 (user reassignment and AV diversions) and Strategy 5 (inclusion of en-route drop-off AVs in AV-user assignment problem) to the base optimization-based assignment strategy, Strategy 3. The AV-user mathematical programming formulation associated with Strategy 6 includes the constraints in Eqn. (21) and Eqn. (22) as well as the additional terms in Eqn. (23) and Eqn. (24).

The right side of Figure 6-6 shows the results of allowing AV diversions (AV *5* is diverted from user *2* to user *3*), user reassignment (user *2* is reassigned from AV *5* to AV *4*), and two-person schedules (AV *1* is assigned to pick up user 4, as it is en-route to drop off user *1*). For this problem instance, the total fleet mileage associated with Strategy 6 is 5.4 miles.

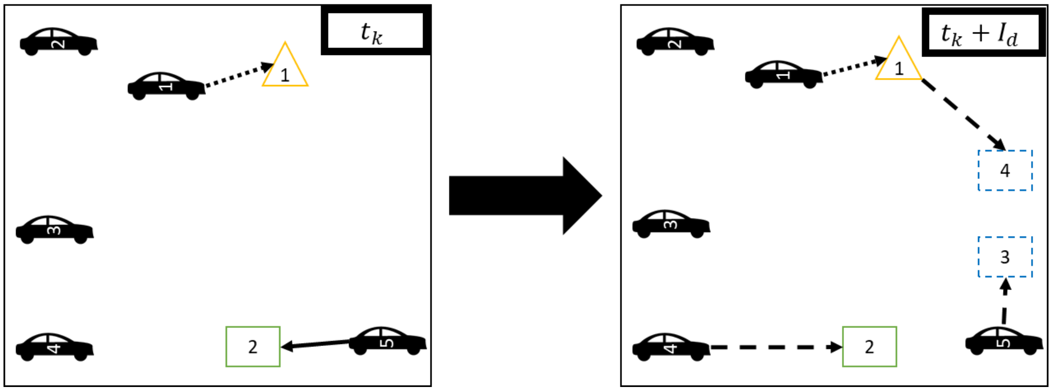


Figure 6-6: AV-user assignment based on Strategy 6 (total fleet miles = 5.4)

Table 6-1 distinguishes between the six AV-user assignment strategies. Section 6.5 compares these six AV-user assignment strategies on much larger problem instances. As the problem is dynamic and stochastic, it is not possible to guarantee any other strategy will perform the best. Hence, a variety of scenarios are presented in the computational results section, to empirically compare the six strategies.

Table 6-1: Overview of AV-User Assignment Strategies

Strategy	FCFS/ Optimization	Users ($\mathcal{C}'$)	AVs ($\mathcal{V}'$)	Sequential/ Simultaneous	User Reassignment?	En-Route Drop-off AV
1	FCFS	C^U	V^I	Sequential	No	No
2	FCFS	C^U	V^I	Sequential	No	No
3	Optimization	C^U	V^I	Simultaneous	No	No
4	Optimization	$C^U \cup C^A$	$V^I \cup V^P$	Simultaneous	Yes	No
5	Optimization	C^U	$V^I \cup V^D$	Simultaneous	No	Yes
6	Optimization	$C^U \cup C^A$	$V^I \cup V^P \cup V^D$	Simultaneous	Yes	Yes

6.5 Experiments and Computational Results

This section presents experiments to compare the efficiency of the AV-user assignment strategies across two metrics. The first metric is average user wait time and the second metric is the ratio of empty fleet miles to total fleet miles, wherein total fleet miles is the sum of empty and loaded fleet miles. The average user wait time metric aims to represent customer service quality. Minimizing wait times should increase the competitiveness of SAMSs with the personal (autonomous) vehicle.

The empty fleet miles metric aims to represent SAMS fleet operational costs. The number of loaded fleet miles is fixed given the transportation system is modeled on a Manhattan grid network with omnipresent streets, and the SAMS does not allow shared rides. Minimizing empty fleet miles should decrease fuel costs, increase the life of the vehicle, and reduce maintenance costs. To compete with the personal vehicle, the SAMS provider can pass these cost savings on to users.

This section presents two sets of computational experiments. In the first set of experiments, the AV-user assignment strategies are compared across thousands (252 x 20) of artificial demand scenarios. In the second set of experiments, the assignment strategies are tested on Chicago taxi data, which represent a realistic spatio-temporal demand pattern.

The experiments aim to compare the six AV-user assignment strategies (across the two metrics) when the fleet size is small relative to the user demand rate. This is likely to happen during the morning and evening peak periods. With a fixed fleet size, efficiently assigning AVs to users can increase the number of user served, decrease average user wait times, and/or reduce operational costs.

Artificial Demand

Experimental Design

Table 6-2 displays the input parameter values for the simulation experiments with artificial demand. The four-hour period represents the morning or evening peak. The fixed fleet size would be most stressed during these two periods.

User requests are generated based on the area size, user demand rate, and spatial demand pattern input parameters. Varying area size and the spatial demand pattern significantly impact trip distance, as shown in Table 6-2. The region sizes are varied to examine the impact of region size, and trip distance on the performance of the AV-user assignment strategies.

In addition to varying the area size and spatial demand pattern, this section presents the performance of the AV-user assignment strategies as a function of the fleet size. Given that SAMSs do not yet exist, and this chapter aims to analyze the case where the fleet size is small relative to the demand rate, the analysis refrains from estimating and selecting a fleet size.

Table 6-2: Parameter Values for Artificial Demand Scenarios

Parameter	Symbol	Value	Units
Simulation Length	t_{max}	4	hours
User Demand Rate		1000	users/hour
Area Size		(a) 16 (b) 64 (c) 256	miles^2
Spatial Demand Pattern		(1) Uniform (2) Clustered	NA
Trip Distance (mean)		(a-1) 2.8 (a2) 3.3 (b-1) 5.4 (b2) 6.7 (c-1) 10.7 (c2) 13.3	miles
Trip Distance (sd.)		(a-1) 1.2 (a-2) 1.1 (b-1) 2.6 (b-2) 2.1 (c-1) 5.2 (c-2) 4.3	miles
Inter-decision Interval	I^d	10	seconds
Vehicle Speed	v	35	mph
Drop-off Time	c_d	15	seconds
Pickup Time	c_p	45	seconds
Weight of Elapsed Wait Time	γ	50	feet/second
En-route drop-off Assgn. Penalty	φ	750 (15)	feet (sec.)
En-route pickup Assgn. Penalty	δ	1500 (30)	feet (sec.)

The combination of six AV-user assignment strategies, seven AV fleet sizes, three area sizes, and two spatial patterns requires 252 (6x7x3x2) unique simulation experiments. Moreover, to produce statistically significant results, each of the 252 simulation experiments were replicated twenty times. A random number generator varies user origins, destinations, and request times across each set of twenty replications.

The simulations were run on a standard 64-bit desktop computer with 8GB of RAM, and a 3.20 GHz processor. A single simulation experiment replication takes anywhere from twenty seconds to ten minutes to complete. Experiments with larger fleet sizes take longer to run, as do experiments with Strategy 4 and Strategy 6, and to a lesser extent Strategy 5.

Uniform Demand Results

Table 6-3 displays the mean and standard error (across twenty replications) of average user wait time for all the uniform synthetic demand scenarios. The standard error is quite low for each scenario because the variance is small across replications and the number of replications is reasonably high.

According to Table 6-3, Strategy 1 is always extremely inefficient. Additionally, Strategy 2, is significantly less efficient than even Strategy 3 when the fleet size is small relative to the demand rate. As fleet size increases, Strategy 2 performs much better than Strategy 1 and similar to the optimization-based strategies. When the fleet size is small relative to the demand rate, Strategy 6 outperforms all the other strategies. However, as the fleet size increases, Strategy 4 outperforms Strategy 6 in terms of average user wait time.

Table 6-3: Average Travel Time (min) Results for Uniform Demand Scenarios

Fleet Size	Strategy 1		Strategy 2		Strategy 3		Strategy 4		Strategy 5		Strategy 6	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Area Size = 16 mi^2												
130	52.4	*0.34*	43.4	*0.34*	10.4	*0.29*	8.8	*0.29*	7.5	*0.25*	6.1	*0.23*
140	45.7	*0.34*	33.8	*0.31*	4.4	*0.21*	3.2	*0.16*	3.2	*0.12*	2.4	*0.09*
150	37.3	*0.30*	25.6	*0.34*	2.5	*0.03*	1.7	*0.04*	1.7	*0.04*	1.5	*0.03*
160	30.1	*0.33*	18.2	*0.35*	1.9	*0.05*	1.2	*0.02*	1.2	*0.02*	1.2	*0.02*
170	23.8	*0.28*	9.1	*1.00*	1.2	*0.03*	1.0	*0.01*	1.0	*0.01*	1.0	*0.01*
175	20.8	*0.27*	4.1	*0.92*	1.1	*0.02*	0.9	*0.01*	1.0	*0.01*	0.9	*0.01*
200	9.0	*0.29*	0.8	*0.01*	0.8	*0.01*	0.8	*0.01*	0.8	*0.01*	0.8	*0.01*
Area Size = 64 mi^2												
230	55.7	*0.29*	47.6	*0.33*	10.5	*0.29*	8.7	*0.30*	8.8	*0.31*	6.8	*0.24*
240	52.3	*0.33*	43.4	*0.34*	7.0	*0.26*	5.4	*0.24*	5.8	*0.19*	4.5	*0.16*
250	49.0	*0.32*	38.3	*0.38*	4.7	*0.14*	3.7	*0.10*	4.1	*0.14*	3.2	*0.09*
255	47.3	*0.31*	36.0	*0.34*	4.3	*0.07*	3.3	*0.07*	3.5	*0.12*	2.9	*0.07*
280	37.6	*0.31*	25.0	*0.33*	3.3	*0.11*	2.0	*0.04*	2.1	*0.04*	2.0	*0.04*
305	28.7	*0.31*	12.6	*1.27*	1.9	*0.05*	1.5	*0.02*	1.7	*0.03*	1.7	*0.02*
330	21.2	*0.30*	2.6	*0.59*	1.5	*0.04*	1.3	*0.02*	1.6	*0.02*	1.6	*0.03*
Area Size = 256 mi^2												
390	64.7	*0.44*	58.6	*0.48*	22.6	*0.46*	20.8	*0.47*	22.0	*0.45*	18.7	*0.43*
400	63.0	*0.43*	56.6	*0.45*	20.4	*0.46*	18.6	*0.48*	20.0	*0.43*	16.8	*0.43*
410	61.1	*0.46*	54.5	*0.49*	18.3	*0.46*	16.5	*0.47*	17.9	*0.46*	14.9	*0.44*
415	60.4	*0.43*	53.5	*0.48*	17.3	*0.45*	15.4	*0.46*	16.9	*0.50*	14.0	*0.44*
440	56.2	*0.45*	48.8	*0.47*	12.7	*0.44*	10.9	*0.44*	12.1	*0.44*	10.1	*0.34*
465	52.1	*0.43*	44.0	*0.51*	9.0	*0.37*	7.4	*0.34*	8.9	*0.32*	7.5	*0.25*
490	48.4	*0.44*	38.9	*0.47*	7.3	*0.18*	5.8	*0.18*	6.7	*0.24*	6.0	*0.15*

The finding that Strategy 4 slightly outperforms Strategy 6, in terms of average user wait time, when the fleet size is high relative to the demand rate is an important one. The results in Table 6-4 indicate that Strategy 6 always outperforms Strategy 4 in terms of empty fleet miles, independent of fleet size and demand rate. This suggests that there is an important trade-off an AV fleet manager may need to consider when choosing between AV-user assignment strategies, during the off-peak period. However, when the fleet size is small relative to the demand rate, the fleet manager should always use Strategy 6.

It is important to highlight the reason why the performance gap in average user wait times across assignment strategies is much greater for small fleets than large fleets. Given the demand rate of user requests is constant across all scenarios, as the fleet size increases, the number of idle vehicles in the service region at any moment in time increases with fleet size. When the fleet size is large, and there are many idle AVs in the service region at time t, less-sophisticated strategies that simply assign idle AVs to users work just fine. Conversely, when the fleet size is small, and there are very few idle AVs in the service region at time t, allowing en-route pickup AVs to be re-assigned, and allowing en-route drop-off AVs to be considered in the assignment problem, in addition to idle AVs, is highly beneficial.

Table 6-4 displays the average (across twenty replications) of the ratio of empty fleet miles to total fleet miles for all the uniform synthetic demand scenarios. The standard error is not presented because it is very small relative to the average value, similar to Table 6-3. Table 6-4 shows that Strategy 1 is terribly inefficient, independent of fleet size. Strategy 2 is also very inefficient when fleet size is low relative to the demand rate, but as the fleet size increases, Strategy 2 approaches the efficiency of the optimization-based strategies. The presentation of assignment strategies suggests why FCFS strategies are inefficient when the fleet size is small relative to the demand rate. If the number of idle AVs is small, and a new user request enters the system, one of the idle AVs is assigned to the new user request, even if the pickup location is very far away.

In terms of empty fleet miles, Strategy 6 unambiguously outperforms all the other strategies across all area sizes and fleet sizes. The size of the performance gap between Strategy 6 and slightly less efficient strategies 4 and 5 increases when the fleet size is small relative to the demand rate. This dominance across experiments suggests that if operational costs are the most important metric for SAMSs, then the fleet manager should employ Strategy 6.

Table 6-4: Ratio of Empty Fleet Miles to Total Fleet Miles Results for Uniform Demand Scenarios

Fleet Size	Strategy 1	Strategy 2	Strategy 3	Strategy 4	Strategy 5	Strategy 6
Area Size = 16 mi^2						
130	49.0%	43.6%	19.8%	18.2%	16.0%	14.5%
140	49.1%	43.3%	20.1%	19.0%	18.2%	16.7%
150	49.0%	42.9%	24.3%	21.5%	18.4%	16.8%
160	49.1%	42.0%	25.4%	19.2%	17.0%	16.0%
170	49.0%	37.5%	20.1%	17.0%	15.8%	15.2%
175	48.6%	27.2%	18.5%	16.2%	15.4%	14.8%
200	48.5%	15.0%	14.8%	14.0%	13.7%	13.4%
Area Size = 64 mi^2						
230	50.2%	45.0%	18.3%	16.7%	16.1%	13.8%
240	50.2%	44.8%	18.5%	17.1%	16.6%	14.8%
250	50.3%	44.5%	19.9%	19.2%	17.4%	15.1%
255	50.2%	44.5%	21.3%	20.2%	17.2%	15.1%
280	50.2%	44.0%	24.0%	18.0%	15.7%	14.4%
305	50.2%	37.6%	18.2%	15.7%	14.6%	13.7%
330	49.6%	20.0%	16.2%	14.6%	13.9%	13.2%
Area Size = 256 mi^2						
390	51.0%	46.6%	17.9%	16.5%	16.1%	13.1%
400	51.0%	46.5%	17.8%	16.5%	16.3%	13.2%
410	51.0%	46.4%	17.9%	16.5%	16.4%	13.3%
415	51.0%	46.3%	17.9%	16.5%	16.4%	13.3%
440	51.0%	45.9%	18.1%	16.8%	16.2%	13.9%
465	50.9%	45.8%	18.9%	18.1%	16.9%	14.6%
490	51.1%	45.5%	22.0%	20.5%	16.8%	14.6%

Clustered Demand Results

This section briefly presents the clustered demand results, which largely follow the same pattern as the uniform demand results. To generate clustered demand, four cluster centroids were placed in the four quadrants of the square service region. Users origins and destinations were each randomly assigned to one of the four quadrants. Their final origin and destination locations are determined by drawing from a random normal distribution centered on their quadrant's cluster centroid, with a standard deviation equal to 5% of the length of an edge of the square service region. In the generation of trip origins and destinations, if the origins and destinations are less than 0.8 miles apart, the user is assigned a new destination.

Table 6-5 displays both the average user wait time and empty fleet miles results for the clustered demand scenarios. The results for Strategy 1 and Strategy 2 are not included because Table 6-3 and Table 6-4 show that these two strategies are significantly less efficient than the optimization-based strategies. Once again, the results indicate that Strategy 6 unambiguously outperforms the other strategies in terms of empty fleet miles across all scenarios. Similarly, when fleet size is small relative to the demand rate, Strategy 6 outperforms the other assignment strategies in terms of average user wait times. However, with large fleet sizes, Strategy 3 and Strategy 4 outperform Strategy 5 and Strategy 6 in terms of average user wait times.

Table 6-5: Computational Results for the Clustered Artifical Demand Scenarios

Fleet Size	Strategy 3	Strategy 4	Strategy 5	Strategy 6		Strategy 3	Strategy 4	Strategy 5	Strategy 6
	Average User Wait Time (min)					Ratio of Empty Miles to Total Miles			
Area Size = 16 mi^2									
130	16.3	15.8	13.8	13.4		11.5%	11.0%	8.2%	7.9%
140	9.4	8.9	7.8	7.4		11.6%	11.1%	8.8%	8.4%
150	3.9	3.4	3.3	3.0		12.2%	11.7%	10.4%	9.7%
160	2.1	1.5	1.8	1.7		16.6%	14.0%	10.5%	9.7%
170	1.5	0.9	1.3	1.3		17.1%	12.7%	10.0%	9.3%
175	1.1	0.8	1.2	1.2		15.2%	11.9%	9.8%	9.2%
200	0.6	0.6	1.0	0.9		10.4%	9.7%	9.0%	8.6%
Area Size = 64 mi^2									
230	21.8	21.1	19.6	19.2		11.3%	10.7%	8.3%	8.0%
240	17.7	17.1	15.9	15.4		11.3%	10.7%	8.6%	8.2%
250	14.0	13.4	12.6	12.1		11.4%	10.8%	8.8%	8.4%
255	12.3	11.7	10.9	10.6		11.4%	10.8%	9.0%	8.6%
280	5.1	4.5	4.8	4.4		12.4%	11.9%	10.4%	9.4%
305	2.9	2.1	2.8	2.8		16.8%	13.2%	10.4%	9.4%
330	1.5	1.3	2.4	2.5		12.8%	11.3%	10.1%	9.2%

Comparing the results in Table 6-5 with those in Table 6-4, the ratio of empty fleet miles is significantly lower in the clustered demand case compared to the uniform demand case. This is unsurprising as AVs are likely to have shorter distances between their drop-off of one user and the pickup of the next user if user origins and destinations are clustered. Nevertheless, this is an important finding, as it suggests, the percentage of empty fleet miles generated by an *on-demand SAMS without shared rides* is likely to be heavily impacted by the spatial demand distribution.

Chicago Taxi Demand

To test the AV-user assignment strategies under a more-realistic spatio-temporal demand distribution, the Chicago taxi data was used (Chicago Data Portal, 2017). The Chicago taxi dataset includes over 100 million taxi trips taken in the Chicago region between 2013 and 2016. A taxi-trip observation includes a user pickup location, drop-off location, and pickup time. In this study, the user's pickup time is treated as the user's request time. Random days from September 2014 were selected from the Chicago taxi data.

Experimental Design

Most of the input parameters for the taxi demand scenarios were the same as those in Table 6-2. However, the taxi simulation is an entire day ($t_{max} = 24\ hours$), the user demand rate (λ) increased to 3000-4000 users per hour on average; however, the user request rate is time-dependent. On Friday and Saturday night the demand rate approaches 10,000-12,000 users per hour. The average trip distance for the taxi trips is surprisingly only around 5 miles; however, the standard deviation of trip distance is around 6 miles. The Chicago service region is approximately 625 mi^2. Lastly, the inter-decision interval (I^d) was increased to 30 seconds.

Taxi Demand Results

Table 6-6 displays the computational results for the taxi demand including the average user wait time and the ratio of empty miles to total miles. In one set of experiments, the full taxi data was used. In several other sets of experiments, 50% of the taxi data was used. The results in Table 6-6 provide more-evidence to support the conclusions in the artificial demand section. Strategy 6 decidedly generates fewer empty miles than the other assignment strategies. Moreover, Strategy 6 outperforms the other strategies in terms of average user wait time when the fleet size is small relative to the demand rate. However, Strategy 4 outperforms Strategy 6 in terms of average user wait time when the fleet size is large relative to the demand rate.

Table 6-6: Experimental Results for Taxi Demand

Fleet Size	Strategy 3	Strategy 4	Strategy 5	Strategy 6		Strategy 3	Strategy 4	Strategy 5	Strategy 6
	Average User Wait Time (min)					Percentage of Empty Miles			
Taxi Day 1 (100% Demand)									
700	5.7	4.9	8.9	7.3		22.7%	21.6%	22.3%	20.4%
750	2.7	2.1	5.6	5.2		23.3%	21.5%	22.6%	19.7%
800	2.2	1.9	4.6	4.6		23.6%	21.6%	21.7%	19.5%
Taxi Day 2 (50% Demand)									
275	23.2	21.2	24.5	19.6		25.6%	24.6%	25.6%	23.1%
300	11.9	10.6	14.2	10.3		27.1%	24.6%	26.4%	23.1%
325	6.6	5.6	8.1	6.1		26.9%	24.7%	26.0%	22.5%
350	3.9	3.2	5.8	5.2		26.6%	24.1%	25.3%	22.5%
375	2.8	2.5	5.1	5.0		26.0%	23.9%	24.7%	22.4%
400	2.8	2.4	5.0	4.8		26.5%	23.6%	24.8%	22.2%
Taxi Day 3 (50% Demand)									
275	8.3	7.0	10.6	6.8		24.8%	22.9%	23.7%	19.8%
300	4.2	3.2	5.5	4.9		26.2%	23.4%	23.2%	19.9%
325	2.7	2.2	4.5	4.4		26.7%	23.3%	23.8%	19.9%
Taxi Day 4 (50% Demand)									
275	9.4	8.3	11.8	8.5		25.3%	24.2%	24.5%	21.3%
300	4.5	3.1	6.3	4.7		27.0%	23.8%	23.7%	20.8%
325	2.6	2.2	4.1	3.7		25.6%	23.5%	23.2%	19.6%

6.6 Conclusion

To reduce operational costs and maximize service quality, it is critical that SAMSs are operated efficiently. As such, this chapter examines an *on-demand SAMS without shared rides* and focuses on modeling and comparing solution strategies for the SAMS's operational problem.

The operational problem is highly-dynamic and stochastic as user requests arrive randomly, users want to be served immediately, and the SAMS fleet controller has no advanced information about the user requests. To solve the problem, the SAMS fleet controller re-solves an AV-user assignment problem in real-time as new requests enter the system. The chapter compares six different AV-user assignment strategies.

The first two assignment strategies are simplistic FCFS assignment strategies; the last four strategies require a mathematical programming solver. The optimization-based strategies, particularly the strategies that involve reassigning (diverting) assigned users (en-route pickup AVs) to other AVs (user requests), significantly outperform the simplistic FCFS assignment strategies. The more-sophisticated optimization-based assignment strategies significantly reduce (empty) SAMS fleet miles and average user wait times when the fleet size is small relative to the demand rate. However, as fleet size increases, the simple assignment strategies are comparable to the more advanced strategies.

SAMS fleet controllers should strongly consider employing optimization-based strategies, strategies that allow en-route pickup AVs to be reassigned, and strategies that incorporate en-route drop-off AVs in the assignment problem. These strategies allow fleet controllers to handle the urgency of new user requests and the stochasticity of future user requests in a computationally-efficient manner.

Another important finding is that the spatial distribution of user requests significantly impacts the percentage of total fleet miles that are empty. Clustered demands result in a lower percentage of empty miles than uniformly distributed demands.

Chapter 7 Assessing the Operational Benefits of Shared-Ride SAMSs

7.1 Overview

This chapter presents an assessment of the operational efficiency benefits, from the perspective of the fleet operator, associated with providing an on-demand *shared-ride* mobility service compared with the *on-demand SAMS without shared rides* in Chapter 6. This chapter defines the *on-demand shared-ride SAMS*, provides a description of the underlying problem associated with operating this SAMS as well as a brief overview of the modeling framework and solution approach for addressing the dynamic stochastic vehicle routing problem.

This chapter focuses on comparing the operational efficiency of the *on-demand shared-ride SAMS* and the *on-demand SAMS without shared rides* in terms of user in-vehicle travel time, user wait time, fleet miles, and fleet size. The results of the analysis clearly show significant operational benefits associated with allowing shared rides. A sensitivity analysis on the maximum detour for user requests indicates that even with a maximum in-vehicle detour distance/time of 5%, the shared-ride SAMS provides meaningful operational benefits. These findings have important implications for SAMS providers, SAMS users, and society as a whole.

7.2 Introduction

Congestion in urban areas is an ongoing and daunting challenge that has many causes (e.g. land-use, auto-dependency, tragedy of the commons/cheap roads) and negative effects (economic inefficiency, stress, harmful vehicle emissions, etc.). There is a lot of excitement surrounding AVs and their potential to address congestion in metropolitan areas. AVs should certainly be able to increase the effective capacity of highway segments (Talebpour and Mahmassani, 2016) and signalized intersections (Levin et al., 2017a). Moreover, if AVs can reduce traffic accidents and incidents that cause lane blockages, this should improve roadway throughput.

Despite these positive aspects of AVs in terms of their ability to address congestion, there are also possible futures wherein AVs make congestion worse. Consider a future where users significantly change their travel behavior, activity behavior, as well as the location of their homes and workplaces due to AVs. If humans no longer have to drive, they can use their in-vehicle time more productively. As such, the disutility associated with sitting in traffic or traveling long distances is quite low if travelers can sleep, watch movies, work, talk on the phone etc. instead of having to drive the vehicle themselves. If they are doing this alone in an AV, and the AV also has to drive empty between dropping of one user and picking up the next, then AVs will significantly increase overall vehicle miles traveled (VMT) and likely increase congestion.

There are many possible paths to avoiding this outcome, and most of these paths involve changes in land-use and transportation policy and planning (e.g. congestion pricing). However, a necessary outcome of the changes in land-use and transport policy and planning must involve a shift away from a transportation system where most travelers travel alone in their own vehicle. There is simply not enough space to add the capacity necessary to reduce congestion if all travelers are using single-occupancy vehicles and zero occupancy AVs. High-capacity transit routes are ideal in terms of moving large numbers of people between certain high-density, high-demand areas. However, there are certain parts of many metropolitan regions that fixed-route transit service does not and cannot provide quality service at a reasonable cost to the transit operator.

In areas where and times-of-day when fixed-route transit service does not provide high-quality service, there is a role for shared-ride mobility services, particularly *on-demand shared-ride SAMSs*. From a user's perspective, using a shared-ride service typically increases in-vehicle travel time while decreasing the cost/price of the service. From a fleet operator's perspective, it is crucial that *on-demand shared-ride SAMSs* are operated efficiently so that users obtain the cost/price savings of shared rides while not experience long in-vehicle travel times or long wait times.

Motivated by the importance of shared-ride mobility services in terms of reducing congestion in the future with AVs, this chapter presents the *on-demand shared-ride SAMS*. The SAMS has the following characteristics:

- Travelers request rides dynamically via a mobile application
- A request includes a pickup location and a drop-off location, both of which must be within a pre-defined geographical service region
- Travelers want to be served (i.e. picked up) immediately
- Travelers will be always be served, assuming they are willing to wait, i.e. the AV fleet operator cannot reject traveler requests
- A single AV picks up and drops off a traveler request i but the same AV may pick up and/or drop off other traveler requests while traveler request i is in the AV
- The AVs in the fleet are functionally homogeneous

- The AVs can only have two traveler requests inside at one time (similar to Lyft Line)
- The AV fleet size is fixed
- The AV fleet operator has complete control over each AV
- The AV fleet operator seeks to minimize fleet miles and traveler wait times; constraints are placed on maximum traveler detours

The operational problem statement and a mathematical model of the *on-demand shared-ride SAMS* are presented briefly in the next section. The focus of this chapter is to compare this *on-demand shared-ride SAMS* with the *on-demand SAMS without shared rides* defined in Chapter 6.2. The comparison is in terms of the operational efficiency of the two on-demand SAMSs. The societal and individual mobility benefits associated with shared-ride mobility services are well-recognized and discussed in the academic and broader literature. However, this section takes the perspective of the SAMS fleet operator and assesses the operational efficiency benefits associated with shared-ride vs non-shared-ride mobility services. The working hypothesis in this study is that the economies of density and scale are much greater in the *on-demand shared-ride SAMS* than the *on-demand SAMS without shared rides* which will allow the *on-demand shared-ride SAMS* to serve more users with the same number of vehicles than the *on-demand SAMS without shared rides* while only increasing users in-vehicle travel time slightly.

7.3 Research Methodology

This section describes the methods employed to compare the *on-demand shared-ride SAMS* and the *on-demand SAMS without shared rides* in terms of operational efficiency.

User Requests

This study uses the NYC taxi data from the borough of Manhattan to represent traveler requests for on-demand SAMSs. Aside from proprietary data owned by mobility service providers, taxi data likely provides the most realistic spatio-temporal distribution of traveler requests for a future SAMS. The taxi trip data includes trip origin, trip destination, and pickup time as well as several other fields not used in this study. The taxi pickup time is treated as the user request time in this study for the on-demand SAMSs.

On-demand SAMS without Shared Rides Fleet Operations

The underlying operational problem associated with the *on-demand SAMS without shared rides* is defined in Section 6.3 of this thesis. The fleet operator needs to assign AVs to user requests as they enter the system dynamically and randomly. In order for a fair comparison between the *on-demand SAMS without shared rides* and the *on-demand shared-ride SAMS*, the solution approach in Chapter 6 is slightly modified in this chapter.

On-demand Shared-ride SAMS Fleet Operations

The *on-demand shared-ride SAMS* operational problem is defined similarly to the *on-demand SAMS without shared rides* in Section 6.3. The operational problem is characterized by a fleet of AVs $\mathcal{V} = \{V_1, V_2, \dots V_j, V_{j+1}, \dots, V_{|\mathcal{V}|}\}$ that aim to serve users $\mathcal{C} = \{C_1, C_2, \dots C_i, C_{i+1} \dots, C_{|\mathcal{C}|}\}$ who request service during the finite time horizon $T = [0, t_{max}]$, over a rectangular geographic service region $\mathcal{G}$ with side lengths L_1 and L_2. The geographical region $\mathcal{G}$ is a Manhattan plane $\mathcal{G} = \{(x, y) | \, x \in [0, L_1], y \in [0, L_2]\}$. The distance between any two locations l_1 and l_2, where $l_1, l_2 \in \mathcal{G}$, is denoted $d(l_1, l_2)$.

At time $t = 0$, AVs may be located at one or several depots, or they may be dispersed throughout the entire region. User requests occur according to an unknown stochastic process $\mathcal{F}^C$. Each user request C_i comes with a request time $t_r^{C_i} \in T$, pickup location $l_p^{C_i} \in \mathcal{G}$, and drop-off location $l_d^{C_i} \in \mathcal{G}$. Let $t_p^{C_i}$ and $t_d^{C_i}$denote the time an AV picks up and drops off user C_i, respectively.

The AV must pick up the user at her requested pickup location and drop her off at her requested location. However, the AVs can pick up an additional user request, even if there is a traveler request currently inside the AV. The goal of the SAMS fleet controller is to efficiently serve the user requests via minimizing user wait times, in-vehicle travel times, and fleet distance.

This study also makes the following modeling assumptions:

- The AVs operate on a Manhattan network with no congestion and no travel time uncertainty
- AV fleet operator has no information (deterministic or stochastic information) about traveler requests until the moment the traveler makes a request via their mobile phone application
- AVs do not have a maximum distance or time constraints (i.e. there is no need to refuel)

This chapter does not detail the stochastic dynamic model for the *on-demand shared-ride SAMS* because it is very similar to the models in Chapter 6 and Chapter 8. However, it does detail the rolling-horizon solution approach employed to solve the stochastic dynamic *on-demand shared-ride SAMS* operational problem.

Let C^o and C^{IV} denote the set of open user requests (meaning, they have not been assigned to an AV yet) and in-vehicle user requests. If t_k is the current time and $t_r^{C_i}$ is the request time of user $C_i \in C^o$, then user C_i's elapsed wait time (w_i) is $w_i = t_k - t_r^{C_i}$. Similarly, let V^I, V^P, and V^D be the set of idle, en-route pickup, and en-route drop-off AVs respectively; $\mathcal{V} = \{V^I, V^P, V^D\}$. Moreover, let V' denote the subset of AVs that are available to be assigned to user requests. In this chapter, V' only include idle and en-route drop-off AVs, not en-route pickup AVs; $V' = \{V^I, V^D\}$.

All idle AVs V^Ican be assigned to all open user requests C^o ; however, in this modeling framework some en-route drop-off AVs V^D are not eligible to be assigned to any open user requests C^o. In other cases, certain en-route drop-off AVs V^D are not eligible to service specific open user requests C^o. Let $d_k^{C_i}$ and $d_{max}^{C_i}$ denote the detour distance of user C_i at epoch k and the maximum detour distance of user C_i. Then, if $d_k^{C_i} \geq d_{max}^{C_i}$, the en-route drop-off AV $V_j \in V^D$ carrying user $C_i \in C^{IV}$ is not allowed to be assigned to another user; these AVs are not considered in the assignment problem at epoch k. Similarly, if assigning en-route drop-off AV $V_j \in V^D$ to an open user request $C_i \in C^o$ would increase the detour distance of either the in-vehicle user inside the AV $d_k^{C_{i^*}}$ or the open user request $d_k^{C_i}$ above their respective maximum detour distances $d_{max}^{C_i^*}$, $d_{max}^{C_i}$, respectively, then the AV-user assignment is not feasible. Let f_{ij} equal one if there is a feasible match between en-route drop-off AV $V_j \in V^D$ and open user request $C_i \in C^o$, and zero otherwise.

At every decision epoch $k \in K$, the SAMS fleet operator solves the mathematical programming problem defined below. The time between epochs is the inter-decision time I^d. Like Chapter 6 and Chapter 8, the math program utilizes the assignment (bi-partite matching) problem structure. The formulation of the myopic AV-user shared-ride assignment problem is given in Eqn. (25)-(29).

$$\min_{x_{ij}} \sum_{i \in C^o} \sum_{j \in V'} x_{ij} \{ c^{VOT} \left(t_{ij}^t + t_{ij}^d - w_i \right) + c^{EDCR} \left(d_{ij} \right) - r^{asgn} \} + c^{share} \sum_{i \in C^o} \sum_{j \in V^{IV}} x_{ij} \tag{25}$$

$$\sum_i x_{ij} \leq 1 \qquad \forall j \tag{26}$$

$$\sum_j x_{ij} \leq 1 \qquad \forall i \tag{27}$$

$$x_{ij} \left(1 - f_{ij} \right) = 0 \qquad \forall i, j \in V^D \tag{28}$$

$$x_{ij} \in \{0,1\} \qquad \forall i, j \tag{29}$$

The objective function includes penalty terms for remaining empty pickup time t_{ij}^t, added user detour time t_{ij}^t, empty distance (d_{ij}) to pick up a user. The parameters c^{VOT}, c^{EDCR} and c^{share} denote the value of time, empty distance cost rate, and the penalty for assigning an open user request to an en-route pickup AV. The objective also includes a reward for assigning an AV to a user (r^{asgn}) and a reward that increases as a function of the elapsed wait time of user i.

The constraint in Eqn. (26) ensures that each AV j is assigned to at most one open user request. The constraint in Eqn. (27) ensures that no more than one AV is assigned to a single open user request. The constraint in Eqn. (28) ensures only feasible AV-user assignments are made.

To model the assignment problem for the *on-demand SAMS without shared rides* only idle AVs are considered in the assignment problem, en-route drop-off AVs are not considered. The assignment problem formulation is otherwise the same.

7.4 Experimental Design

This section presents the experiments designed to compare the operational efficiency of the *on-demand SAMS without shared rides* against the *on-demand shared-ride SAMS* using several metrics, including in-vehicle travel time (IVTT), wait time, and vehicle miles travelled (VMT). The IVTT associated with an *on-demand shared-ride SAMSs* is guaranteed to increase relative to an *on-demand SAMSs without shared rides*. However, the experiments aim to determine the relative trade-off between increases in IVTT and potential decreases in wait time and VMT associated with a shared-ride SAMS.

To compare the operational efficiency of the *on-demand SAMS without shared rides* against the *on-demand shared-ride SAMS*, the following parameter values vary across scenarios:

- Demand level: 15%, 20%, 25%, 30%, 35%, and 40% (of Manhattan, NY taxi demand requests)

- Maximum user detour distance/time: 0%, 5%, 10%, 15%, 30%, 50%, 80% (increase over a user's shortest path distance/time)

Moreover, 10 different days of Manhattan, NY taxi demand are treated as 10 independent sets of input data.

Table 7-1 displays the input parameters that do not vary across the experiments. Previous chapters in this thesis discuss reasons and rationalizations for these parameter values. The only new parameter is the 'shared ride penalty' parameter. In this set of experiments, the shared-ride penalty is set to zero in order to not penalize or reward the assignment of users to en-route drop-off AVs compared to idle AVs.

Table 7-1: Parameter Values for Experiments

Parameter	Value	Value	Units
Simulation Length		21.5	hours
Fleet Size		400	vehicles
Inter-decision Interval	I^d	15	seconds
Vehicle Speed		5	meters / second
Drop-off Time		15	seconds
Pickup Time		45	seconds
Shared Ride Penalty	c^{share}	0.0	$
Value of Time	c^{VOT}	23	$/hour
Cost Rate	c^{EDCR}	0.50	$/mile
Assignment Reward	r^{asgn}	5.0	$

7.5 Results

This section presents the results of the computational experiments described in the previous section. In Figure 7-1 through Figure 7-4 the x-axis is the demand level measured in terms of the percentage of total taxi demand in Manhattan, NY on a given day. On average across ten days, 15% of the taxi demand represents around 14,200 user requests and 40% of the taxi demand represents around 37,900 user requests. The different color lines in Figure 7-1 through Figure 7-4 represent different maximum user in-vehicle detour distance/time constraints. The blue line denoted '1' in the legend, represents the *on-demand SAMS without shared rides*. Conversely, the other lines represent different cases of the *on-demand shared-ride SAMS*. For example, the orange line, denoted '1.05' in the legend, is an *on-demand shared-ride SAMS* where a constraint is added such that a user's maximum in-vehicle detour time/distance is 5%.

Figure 7-1 displays the average user in-vehicle travel time results. The results show that relaxing the constraint on maximum detour time/distance increases average user in-vehicle travel time. Moreover, the impact increases as the demand level increases. However, there are couple important things to note. First, even at the highest demand levels and largest user detour time/distance constraint (80%), the increase in average in-vehicle travel time between the *on-demand SAMS without shared rides* and the *on-demand shared-ride SAMS* is still just 2 minutes or a 20% increase. Second, the results in Figure 7-1 should be considered within the context of the results in Figure 7-2 and Figure 7-3.

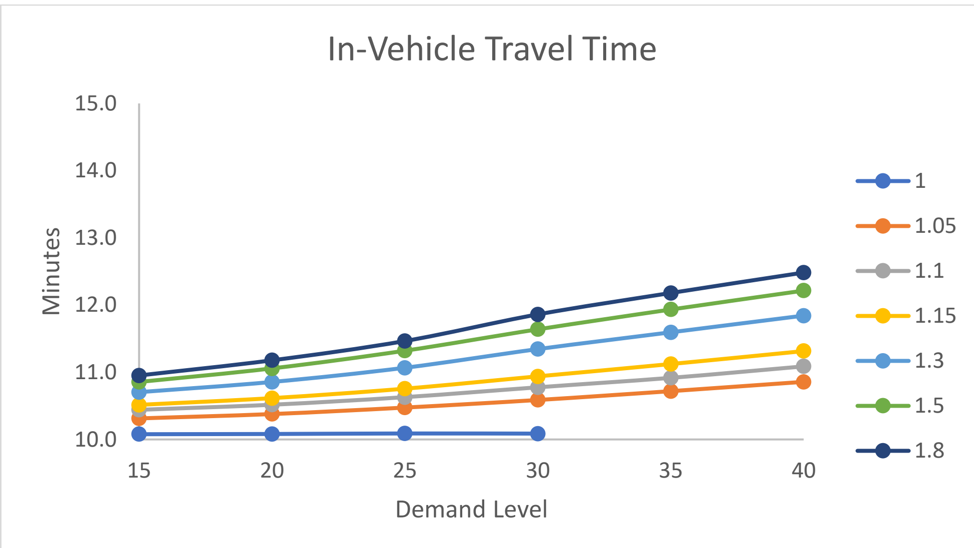


Figure 7-1: Average user in-vehicle travel time as a function of demand (x-axis) and maximum user detour time/distance (line color)

Figure 7-2 displays the average user wait time results. These results clearly illustrate the significant operational benefits of an *on-demand shared-ride SAMS* relative to an *on-demand SAMS without shared rides*. The blue line denoting the *on-demand SAMS without shared rides* shows that at demand levels greater than 25%, the average user wait time begins to increase exponentially with demand level. The *on-demand SAMS without shared rides* is effectively unable to serve demand levels greater than 25% or 30%. Conversely, the *on-demand shared-ride SAMS* results with different maximum detour distance/time constraints appear to be able to handle demand levels of 35-40% with the same fleet size. Even in the case where the maximum detour distance/time is just five percent, the *on-demand shared-ride SAMS* is able to serve a significantly higher volume of demand than the *on-demand SAMS without shared rides*.

This finding suggests that if SAMS providers can incentivize users to opt-in to shared-ride services, they can serve these users with a minimal increase in in-vehicle travel with a much smaller fleet of vehicles than is needed for an SAMS without shared rides.

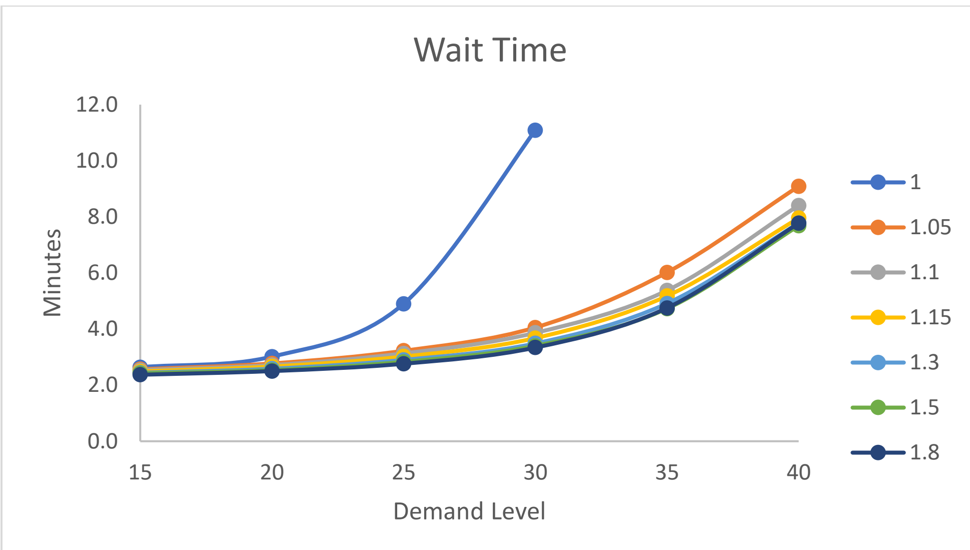


Figure 7-2: Average user wait time as a function of demand (x-axis) and maximum user detour time/distance (line color)

Figure 7-3 displays the total fleet vehicle miles traveled (VMT) results. Although not as dramatic as the average user wait time results in Figure 7-2, the fleet VMT results in Figure 7-3 do illustrate the operational efficiency benefits of an *on-demand shared-ride SAMS* relative to an *on-demand SAMS without shared rides*. The blue line for the *on-demand SAMS without shared rides* shows a near-linear relationship between demand level and fleet VMT. There are economies of scale and density for the *on-demand SAMS without shared rides*; however, they are not very strong. Conversely, the other lines in Figure 7-3 show that for an *on-demand shared-ride SAMS* the relationship between demand level and fleet VMT is sublinear. This is due to the noticeable economies of scale and density associated with an *on-demand shared-ride SAMS*. As demand levels increase, the number of shared-ride opportunities increases, allowing the fleet operator to serve more demand per fleet mile.

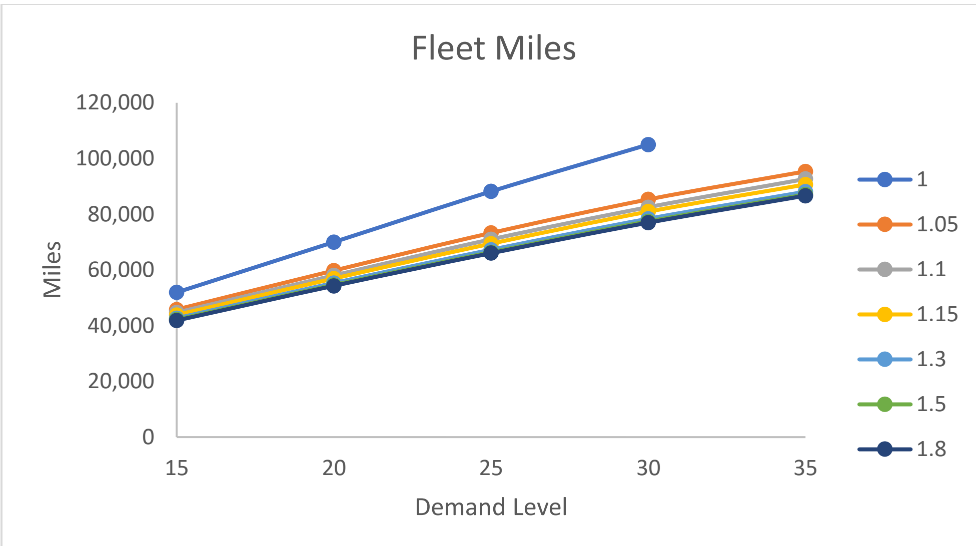


Figure 7-3: Total fleet miles as a function of demand (x-axis) and maximum user detour time/distance (line color)

Figure 7-4 displays the relationship between the percentage of users that share rides and the demand level. The results clearly indicate that as the demand rate increases, the percentage of travelers sharing a ride increases. This makes sense intuitively because as the demand rate increases the number shared-ride opportunities increases. However, it is crucial to understand that the results in Figure 7-4 illustrate why the *on-demand shared-ride SAMS* provides significant operational benefits relative to the *on-demand SAMSs without shared rides*. Figure 7-4 essentially shows that an *on-demand shared-ride SAMS* can increase its effective service rate as the demand rate increases; whereas, the *on-demand SAMSs without shared rides* cannot.

The operational efficiency implications of the findings in Figure 7-2, Figure 7-3 and Figure 7-4 are as follows, relative to an *on-demand SAMSs without shared rides*, the *on-demand shared-ride SAMS* can:

- Serve the same number of travelers with a much smaller fleet size, thereby reducing capital costs;
- Serve the same number of travelers with significantly fewer fleet miles, thereby reducing operational costs;
- Handle surges in demand much more effectively, thereby increasing the resiliency of the SAMS.

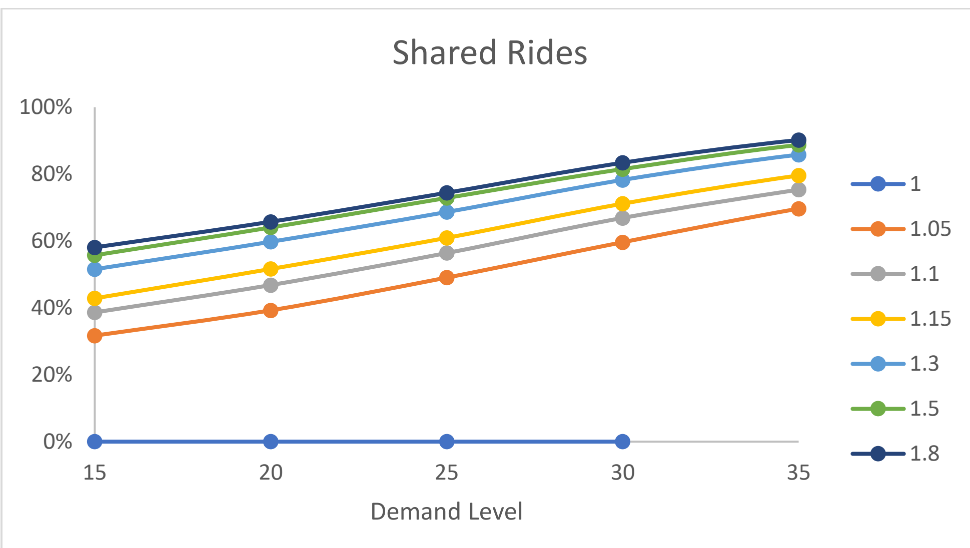


Figure 7-4: Percentage of users who shared a ride as a function of demand (x-axis) and maximum user detour time/distance (line color)

7.6 Conclusion

This chapter presents the *on-demand shared-ride SAMS* and its underlying operational problem. Then it compares the *on-demand shared-ride SAMS* with the *on-demand SAMS without* *shared rides* in terms of operational efficiency. The societal benefits of higher occupancy vehicle travel include lower overall VMT and potentially a reduction in congestion, fuel consumption, and vehicle emissions. The individual mobility benefits of shared rides include a travel option that is cheaper than one person per mobility service vehicle. These advantages are relatively well documented. However, there do not appear to be in-depth analyses examining the operational efficiency advantages of shared-ride mobility service.

The analysis presented in this chapter fills that gap. The results indicate significant operational advantages of an *on-demand shared-ride SAMS* over an *on-demand SAMS without shared rides*. The advantages appear to stem from the economies of density and scale that come with offering shared rides. As demand increases, the number of shared-ride opportunities naturally increases. The *on-demand shared-ride SAMS* can capture these benefits while the *on-demand SAMS without shared rides* cannot. Hence, given the same fleet size the *on-demand shared-ride SAMS* can serve significantly more travelers, using fewer fleet miles, and more effectively handle unexpected surges in demand.

The findings in this chapter suggest that from a purely operational perspective, SAMS providers can benefit significantly from offering and incentivizing users to opt-in to shared-ride services. Of course, there are practical challenges associated with shared-ride services, including, having travelers switch from a personal vehicle to sharing vehicles with strangers. There are also issues that occur when vehicles cannot find travelers on the side of the road, or travelers do not come right away when the vehicle arrives, that need to be addressed. Nevertheless, the findings in this chapter illustrate that if a mobility service provider can address these issues there are significant advantages to shared-ride services compared to mobility services without shared rides.

Chapter 8 On-demand Autonomous Carsharing Service

8.1 Overview

This chapter presents an *on-demand autonomous carsharing service (ODACS)* and defines the underlying operational problem associated with this autonomous mobility-on-demand (AMOD) service where users reserve fully-autonomous vehicles for a user-specified time-slot. An ODACS combines the benefits of existing MOD services (i.e. on-demand service and at-origin pickups) with the benefits of existing carsharing services (i.e. the ability to make multiple trips and temporarily store items).

The ODACS operational problem is a stochastic dynamic control problem (SDCP) where the fleet operator only has probabilistic information about the spatio-temporal distribution of future user requests. This chapter presents a framework to model the ODACS operational problem that captures the evolution of the stochastic dynamic system and the sequential nature of decisions. At each decision epoch, AVs can be assigned to open user requests, repositioned to different subregions of the service area, or remain at their current location. The study presents an operational policy that involves solving a joint AV-user assignment and empty AV repositioning decision problem. The objective function of the joint assignment-repositioning problem captures immediate costs (unproductive vehicle kilometers and user wait times) and immediate rewards (serving open user requests) and approximates costs-to-go through an expected subregion supply-demand imbalance term.

Results indicate that the joint assignment-repositioning operational policy significantly outperforms optimization-based assignment-only policies and a nearest neighbor assignment policy in terms of average user wait times and system cost. The system cost metric is a weighted combination of user wait times and unproductive fleet kilometers.

8.2 Motivation

Mobility-on-demand (MOD) services provide users door-to-door transportation service on-demand, without requiring users to purchase, insure, maintain, or park a personal vehicle, and without having to walk to or from a transit stop/station or wait outside for a transit vehicle. These attractive features suggest why MOD services, like the services offered by ridesourcing companies Uber and Lyft, have seen significant growth and obtained a sizable market share over the past five years in certain areas (Clewlow and Mishra, 2017).

Existing carsharing services allow users to reserve a vehicle for a user-specified time slot (e.g. 30 minutes or three hours). However, unlike MOD services, carsharing services do not offer on-demand pickup service, and users need to access (typically via walking) a carsharing station or carsharing vehicle before starting their carsharing trip. These unattractive features of existing carsharing services, along with several others listed below, suggest why, despite being around for a longer time, carsharing services have not grown as quickly as ridesourcing services throughout the United States. Fortunately, fully-autonomous vehicles (AVs) should allow mobility service providers to offer an on-demand autonomous carsharing service (ODACS) that combines the benefits of MOD services and existing carsharing services.

The main advantage of carsharing services over existing MOD services, for some users and trip purposes, is the ability of carsharing users to reserve a vehicle for a user-specified time slot. Within the time slot, the carsharing user can make multiple trips, collect other users, and/or store items in the vehicle. Examples of this include picking up children from multiple schools, dropping children off at multiple after-school activities, shopping at multiple stores, storing items in the vehicle between stores, stopping to pick up a few things on the way to/from work, and/or traveling to slightly remote regions – such as a park in a rural area – for a few hours. Existing point-to-point MOD services either cannot or struggle to serve these users/trip purposes. Given the average round-trip commute trip chain [7]in the U.S. includes 0.43 non-work stops (Wang, 2015), and non-commute trip chains include significantly more stops per trip chain, there is likely to be significant demand for a MOD service that allows AV reservations for a user-specified period.

There are several negative aspects of existing carsharing services that AVs can overcome. First, as mentioned previously, carsharing users need to walk to a carsharing vehicle (in free-floating carsharing systems) or carsharing station (in roundtrip or one-way carsharing systems) before making a carsharing trip. Second, carsharing service providers either (i) need to hire drivers to rebalance (reposition) vehicles between stations (within a service region) in a one-way (free-floating) carsharing system or (ii) suffer from severe supply-demand imbalances at carsharing stations within the service region. Third, given the need to access a carsharing vehicle or station

[7] A trip chain is a series of trips that start and end at the same location. For example, a home-to-work trip followed by a work-to-home trip is a home-based trip chain. A work-to-lunch, lunch-to-pharmacy, pharmacy-to-work series of trips is a work-based trip chain.

via walking, users often cannot reserve or access a carsharing vehicle within a reasonable distance when they want to travel; hence, individuals cannot rely on carsharing services as a primary mode of transportation. Fortunately, AVs overcome each of these negative aspects since AVs can travel empty to pick up users and reposition themselves to serve future user requests.

The carsharing company Car2Go recently put out a white paper identifying five components of futuristic autonomous carsharing service: fleet management, demand prediction, real-time control of AV fleets (or as they call it ‘fleet intelligence’), intelligent charging, and user experience (Car2go, 2017). This study focuses exclusively on the real-time control of AV fleets with the assumption that the fleet operator has high-quality short-term demand predictions.

Motivated by the ability of AVs to combine the benefits of existing MOD services and carsharing services, this chapter presents, defines, mathematically models, and introduces an operational policy for the ODACS operational problem. This chapter defines the ODACS as having the following characteristics:

- Users request a pickup location, drop-off location, and usage time/reservation length via a mobile phone application;
- Users want AVs to come to their pickup location immediately;
- Users have complete control over the AV until they release the AVs back to the fleet controller;
- AVs in the fleet are functionally homogenous;
- The service provider owns and operates its own fleet of AVs;

- The fleet controller will not reject user requests within the pre-defined service region.

The ODACS operational problem is a stochastic dynamic control problem (SDCP). The problem is dynamic because decisions need to be made sequentially as new user requests enter the system. The model in this chapter includes two forms of stochasticity, namely, future user requests, and the actual AV usage times for each user. The ODACS operational problem in this chapter is not significantly different from the *on-demand SAMS without shared rides* operational problem in Chapter 6. The main difference is that in the ODACS problem, the SAMS fleet operator cannot easily determine when each in-use AV will be done serving its current passenger. In both cases, an in-use AV cannot be used to serve new traveler requests. However, in the *on-demand SAMS without shared rides* problem, the fleet operator has complete control over the AVs plans and arrival times at each vehicle stop.

The goal of the ODACS operational problem is to minimize a weighted combination of user wait times and empty AV kilometers. Short user wait times are likely to be a key factor in the success of AMOD services (Krueger et al., 2016), especially if users are to forego owning their own vehicles and rely on AMOD services for all travel. Minimizing empty fleet distances is always an important objective as it correlates heavily with operational costs and profitability.

To solve SDCPs, the goal is to find an optimal policy/strategy. A strategy provides a decision function that returns an explicit decision for every possible system state. A simplistic operational strategy for the ODACS operational problem involves immediately assigning each new user request to the closest empty AV. More effective strategies include optimization-based bipartite matching of AVs and users, repositioning AVs proactively, queuing user requests temporarily before assigning them to AVs, and combinations of these policy elements.

This chapter presents an optimization-based joint AV-user assignment and empty AV repositioning operational strategy. The joint assignment-repositioning formulation captures the immediate rewards, immediate costs, and costs-to-go given the system state and decision at each decision epoch. The chapter compares this operational strategy to myopic optimization-based assignment-only strategies, and a strategy that immediately assigns new user requests to the nearest empty AV.

This chapter makes several scientific contributions to the existing literature. First, it formalizes the ODACS operational problem using a modeling framework that captures the stochastic and dynamic elements of the problem. Hanna et al. (2016) and Dandl and Bogenberger (2018a) also study and conceptually define autonomous carsharing services; however, as far as the author is aware, this is the first study to formalize the ODACS operational problem using a modeling framework that captures the stochastic and dynamic elements of the problem. Second, the chapter presents a flexible optimization-based operational strategy that models the sequential decision problem as a joint AV-user assignment and empty AV repositioning problem. Third, both the decision problem's objective function and the analysis' performance metric combine fleet operational costs and user wait time costs into a total system cost metric with monetary units.

8.3 ODACS Problem

This section presents a formal description of the ODACS operational problem and then presents of model of the stochastic dynamic problem.

Problem Statement

The ODACS operational problem is characterized by a fleet of AVs $\mathcal{V} = \{V_1, V_2, \dots V_j, V_{j+1}, \dots, V_{|\mathcal{V}|}\}$ that aim to serve users $\mathcal{C} = \{C_1, C_2, \dots C_i, C_{i+1} \dots, C_{|\mathcal{C}|}\}$ who request an AV rental during the finite time horizon $T = [0, t_{max}]$, over a rectangular geographic service region $\mathcal{G}$ with side lengths L_1 and L_2. The geographical region $\mathcal{G}$ is a Manhattan plane $\mathcal{G} = \{(x, y) | \, x \in [0, L_1], y \in [0, L_2]\}$. The region $\mathcal{G}$ can be divided up into a set of $|\mathcal{Z}|$ equally-sized rectangular subregions (i.e. zones) $\mathcal{Z} = \{Z_1, Z_2, \dots Z_h, Z_{h+1} \dots, Z_{|\mathcal{Z}|}\}$ each with a demand-weighted centroid location $l^{Z_h} \in \mathcal{G}$.

The distance between any two locations l_1 and l_2, where $l_1, l_2 \in \mathcal{G}$, is denoted $d(l_1, l_2)$. Given the AVs travel at fixed vehicle speed v, the travel time between points $t^t(l, l_2)$ is the ratio of $d(l_1, l_2)/v$.

At time $t = 0$, AVs may be located at one or several depots, or they may be dispersed throughout the entire region. User requests occur according to a known stochastic process $\mathcal{F}^C$. Each realized user request C_i is associated with a request time $t_r^{C_i} \in T$, pickup location $l_p^{C_i} \in \mathcal{G}$, drop-off location $l_d^{C_i} \in \mathcal{G}$, and requested usage time $t_{u^r}^{C_i}$, which differs from the user's actual usage time $t_{u^a}^{C_i}$. The AV must pick up the user at her requested pickup location. Let $t_p^{C_i}$ denote the time an AV picks up user C_i. After pickup, the user is free to use the AV as she chooses until returning the AV to her drop-off location. User C_i's actual usage time $t_{u^a}^{C_i}$ is a random variable defined as follows:

$$t_{u^a}^{C_i} = t_{u^r}^{C_i} - N(\mu_u, \sigma_u)$$

Hence, the fleet controller only has probabilistic information on the user's remaining usage time.

The goal of the AV fleet is to efficiently serve the user requests via minimizing user wait times and empty fleet kilometers. Letting $d_e^{V_j}$ denote the cumulative empty distance of AV V_j, then the fleet controller aims to minimize the expected total cost:

$$\min \mathbb{E}\left[c^{VOT}\sum_{C_i\in\mathcal{C}}\left(t_p^{C_i}-t_r^{C_i}\right)+c^{EDCR}\sum_{\mathrm{V_j}\in\mathcal{V}}d_e^{V_j}\right]$$

where c^{VOT} denotes the value of user wait time [$/time] and c^{EDCR} denotes the empty distance cost rate [$/distance].

To meet this objective, the fleet controller has several decision levers. As new user requests enter the system, the fleet controller needs to assign AVs to the requests. However, the fleet controller can position AVs, based on short-term demand forecasts, so that idle AVs are already near new user requests as they enter the system. Specifically, at each decision epoch, AVs can (i) be assigned to open user requests, (2) be repositioned to other subregions, or (iii) stay in their current location.

Model

This chapter makes the following modeling assumptions:

- The carsharing fleet size is fixed in the short term (i.e. one-day)
- The AVs operate on a Manhattan-plane with a fixed vehicle speed v
- Users will wait indefinitely to be served
- The fleet controller only has a stochastic spatio-temporal distribution of future user requests

- The fleet controller only has a stochastic distribution on the actual usage time of users

Given these assumptions, and the problem statement presented above, this section presents a model of the ODACS operational problem.

Decision Epochs

This chapter assumes a finite horizon with a pre-defined number of decision epochs. Let K be the set of decision epochs, |K| the number of decision epochs, and k the index of the k$^{\text{th}}$ decision epoch. The time between decision epochs is a fixed value, denoted I^d and referred to as the inter-decision time. The variable $t_k \in T$ denotes the time of decision epoch $k \in K$.

States

The state variable S_k contains all the information necessary to model the problem from the current epoch $k \in K$, to the end of the modeling period $|K| \in K$.

For the ODACS, the system state at time k (S_k) includes several dimensions. This chapter delineates three sets of entities – users, AVs, and subregions – with states that need to be updated. Let the states of users, AVs, and subregions at epoch $k \in K$ be denoted S_k^C, S_k^V, and S_k^R respectively, where S_k is completely defined by the set $(S_k^C, S_k^V, S_k^Z\ , \mathrm{t_k})$.

User State. The user state S_k^C is the tuple $\left(\sigma_k^{\mathcal{C}}, w_k^{\mathcal{C}}, \mathcal{C}\right)$ denoting the status, elapsed wait time, and static information of users, respectively. The static user information includes user request times $t_r^{\mathcal{C}}$, user pickup locations $l_p^{\mathcal{C}}$, user drop-off locations $l_d^{\mathcal{C}}$, and requested usage times $t_u^{\mathcal{C}}$.

For each user C_i, user status $\sigma_k^{C_i}$ takes on a value in the set {0,1,2,3,4}:

$$\sigma_k^{C_i} = \begin{cases} 0, & C_i \text{ has not requested service by time } t_k \\ 1, & C_i \text{ has requested service but has not been assigned by } t_k \\ 2, & C_i \text{ has been assigned but not picked up by } t_k \\ 3, & C_i \text{ has been picked up and is using the AV at } t_k \\ 4 & C_i \text{ has been served by } t_k \end{cases}$$

If user C_i has not requested service at t_k (i.e. $\sigma_k^{C_i} = 0$), then the static information associated with user C_i is unknown to the fleet controller. The elapsed wait time of a user $w_k^{C_i}$ is simply the difference between the current time and the user's request time: $w_k^{C_i} = t_k - t_r^{C_i}$.

AV State. The AV state $S_k^{\mathcal{V}}$ is the tuple $\left(\varphi_k^{\mathcal{V}}, l_k^{\mathcal{V}}, r_k^{\mathcal{V}}, a_k^{\mathcal{V}}\right)$ denote the status, location, route, and planned arrival times of every AV V_j at decision epoch k, respectively. For each AV V_j, $\varphi_k^{V_j}$ takes on a value in the set {0,1,2,3}:

$$\varphi_k^{V_j} = \begin{cases} 0, & V_j \text{ is idle at time } t_k \\ 1, & V_j \text{ is en-route to pick up a user at } t_k \\ 2, & V_j \text{ is in-use at } t_k \\ 3, & V_j \text{ is repositioning at } t_k \end{cases}$$

The operational strategies presented in this chapter vary the AVs included in the assignment-repositioning decision problem, at decision epoch k, based on their statuses $\varphi_k^{V_j}$.

The set of AV routes $r_k^{\mathcal{V}} = \left(r_k^{V_1}, r_k^{V_2}, \dots, r_k^{V_j}, \dots, r_k^{V_{|\mathcal{V}|}}\right)$ provides the sequenced set of locations AVs will visit next, at decision epoch k. In this chapter, an AV route $r_k^{V_j}$ can only include a maximum of two locations.

The planned arrival times for all AVs $a_k^{\mathcal{V}} = \left(a_k^{V_1}, a_k^{V_2}, \dots, a_k^{V_j}, \dots, a_k^{V_{|\mathcal{V}|}}\right)$ at decision epoch k are probabilistic due to the uncertainty in users actual usage times. The decision maker needs a method to estimate each in-use user's remaining usage time. Using known information such as each user's requested usage time $t_{u^r}^{C_i}$ and the parameters μ_u and σ_u, it is possible to estimate each user's actual usage time as: $t_{u^e}^{C_i} = t_{u^r}^{C_i} - b_u$, where b_u represents a buffer on user's actual usage time.

Subregion State. The subregion state S_k^Z is denoted by $\left(I_k^Z\right)$, the expected imbalance between supply and demand in subregion Z_h between the time t_k at decision epoch k, and the end of the prediction horizon $t_k + h^p$, where h^p is the length of the prediction horizon.

The expected imbalance ($I_k^{Z_h}$) in subregion Z_h is the difference between the expected supply and expected demand in subregion Z_h over the prediction horizon h^p. The expected supply in Z_h is the summation of:

- the number of idle and repositioning AVs currently in subregion Z_h; and
- the number of en-route pickup and in-use AVs that are expected to terminate and become idle in subregion Z_h within the prediction horizon h^p.

Note that the currently repositioning AVs traveling towards subregion Z_h are not included in the expected supply of Z_h. This is because they are counted in the region in which they are currently located at epoch k. The decision problem does not differentiate between currently idle and repositioning AVs at each decision epoch.

The expected demand in Z_h is the summation of:

- the number of open user requests currently in subregion Z_h; and

- the expected number of future requests in subregion Z_h over the prediction horizon h^p (this value comes from a demand forecasting model).

This chapter assumes high-quality spatio-temporal demand forecasts are available at a spatial resolution of 0.45 km^2 and temporal resolution of 5-min. Using this spatial aggregation, the number of subregions $|\mathcal{Z}|$ in Manhattan is 256.

Decisions

At each decision epoch k, given the state of the system S_k, the AV fleet can control the system via changing the plans of AVs. Let X_k denote the set of decision variables at decision epoch k. To model the decision problem, this chapter introduces two variables, x_k^{ji} and y_k^{jh}, defined as follows:

$$x_k^{ji} = \begin{cases} 1, & \text{if AV } V_j \text{ is assigned to pick up user } C_i \text{ at time } t_k \\ 0, & \text{otherwise} \end{cases}$$

$$y_k^{jh} = \begin{cases} 1, & \text{if AV } V_j \text{ is assigned to reposition to zone } Z_h \text{ at time } t_k \\ 0, & \text{otherwise} \end{cases}$$

There are a couple constraints on the decision variables, displayed in Eqn. (30)-(33)

$$\sum_i x_k^{ji} + \sum_h y_k^{jh} \leq 1 \qquad \forall j, h, k \tag{30}$$

$$\sum_j x_k^{ji} \leq 1 \qquad \forall i, k \tag{31}$$

$$x_k^{ji} \in \{0,1\} \qquad \forall i, j, k \tag{32}$$

$$y_k^{jh} \in \{0,1\} \qquad \forall j, h, k \tag{33}$$

The constraint in Eqn. (30) ensures each AV V^j is assigned to at most one open user request C^i or subregion Z^h. The constraint in Eqn. (31) ensures that no more than one AV is assigned to an open user request C^i. The constraints in Eqn. (32) and (33) ensure the two sets of decision variables take on binary values. Also, the constraint below ensures that only 'available' users, denoted $\mathcal{C}'$, and 'available' AVs, denoted $\mathcal{V}'$, are assigned. The users and AVs that are considered 'available' change based on the operational strategy

$$x_k^{ij} + y_k^{jh} = 0 \qquad\qquad \forall i \notin \mathcal{C}', \forall j \notin \mathcal{V}'$$

Exogenous Information

The ODACS operational problem includes two sources of stochasticity in which exogenous information enters the system, namely, the user requests themselves and the actual usage times. The exogenous information that enters the system between decision epochs $k-1$ and k is denoted ω_k. Hence, let $\omega_{k+1} = \{\gamma_k, \delta_k\}$, where γ_k is the set of previously unrequested user requests ($\sigma_k^C = 0$) with a request time between t_k and t_{k+1}; i.e. $\gamma_k \subseteq \left\{C_i | \sigma_k^{C_i} = 0, t_k < t_r^{C_i} \le t_{k+1}\right\}$.

Similarly, δ_k is the set of served user requests that officially release their AVs between t_k and t_{k+1}; i.e. $\delta_k \subseteq \left\{C_i | \sigma_k^{C_i} = 3, t_k \le t_d^{C_i} \le t_{k+1}\right\}$, where user C_i's drop-off time $t_d^{C_i}$ is the summation of her pickup time and actual usage time $t_d^{C_i} = t_p^{C_i} + t_{u^a}^{C_i}$.

Transition Function

The transition function defines how the state of the system S_k updates from decision epoch k to the next decision epoch $k+1$. The decision epoch time t_k updates as follows:

$$t_{k+1} = t_k + I^d$$

The user state S_k^C contains two elements $\left(\sigma_k^{\mathcal{C}}, w_k^{\mathcal{C}}, \mathcal{C}\right)$. The user information ($\mathcal{C}$) stays the same; whereas the users' statuses ($\sigma_k^{\mathcal{C}}$) updates as follows:

$$\sigma_{k+1}^{C_i} = \sigma_k^{C_i} + \mathbb{1}_{C_i \in \gamma_k} + \sum_j x_k^{ji} + \mathbb{1}_{\boldsymbol{t_k} \leq t_p^{C_i} < t_{k+1}} + \mathbb{1}_{\boldsymbol{t_k} \leq t_d^{C_i} < t_{k+1}}$$

Where $\mathbb{1}_{C_i \in \gamma_k}$ equals 1 if user C_i requests service between t_k and t_{k+1}; $\sum_{\boldsymbol{j}} x_k^{ji}$ denotes if user C_i is assigned to an AV; and $\mathbb{1}_{\boldsymbol{t_k} \leq t_p^{C_i} < t_{k+1}}$ and $\mathbb{1}_{\boldsymbol{t_k} \leq t_d^{C_i} < t_{k+1}}$ equal 1 if user C_i is picked up or dropped off by an AV, respectively. Updating user elapsed wait times ($w_k^{\mathcal{C}}$) is straightforward, $w_{k+1}^{C_i} = w_k^{C_i} + I^d$.

The AV state S_k^V contains four elements $\left(\varphi_k^{\mathcal{V}}, l_k^{\mathcal{V}}, r_k^{\mathcal{V}}, a_k^{\mathcal{V}}\right)$. Because the notation is rather cumbersome for the transition function, despite the algebraic relationships being quite simple, this thesis does not display the full transition functions. Updating each AV's status, $\varphi_k^{\mathcal{V}}$, involves checking if the AV:

- reached its pickup location: $\mathbb{1}_{l_k^{V_j} = l_p^{C_i}}$
- was released by the user ($\mathbb{1}_{\boldsymbol{t_k} \leq t_d^{C_i} < t_{k+1}}$)
- was assigned to pick up a user: $x_k^{ji} = 1$
- or was assigned to a new subregion: $y_k^{ji} = 1$

These events also impact the planned vehicle routes $r_k^{\mathcal{V}}$, and the planned arrival times at the stops along the route, $a_k^{\mathcal{V}}$. Finally, updating each AV's location, $l_k^{\mathcal{V}}$, involves moving en-route pickup and repositioning AVs one step towards their pickup and repositioning location, respectively. It is also necessary to check if the AV was released by the user.

The subregion state $S_k^R = \left(I_k^Z\right)$ is recomputed at every decision epoch k using the method of determining the expected supply and expected demand over the prediction horizon h_p as outlined in the States subsection above.

Objective Function

Let $C(S_k, X_k)$ denote the cost of being in state S_k and making decision X_k. For SDCPs, the solution is a policy $\pi \in \Pi$. Each policy π maps states to decisions; i.e. given S_k, policy $\pi \in \Pi$ yields decision $X_k^\pi(S_k)$. The objective of an SDCP is to determine an optimal policy $\pi^* \in \Pi$ that minimizes the objective function in Eqn. (34), subject to the constraints in Eqn. (30)-(33) on the decision variables.

$$\min_{\pi \in \Pi} E^\pi \left[\sum_{k \in \mathrm{K}} C\left(S_k, X_k^\pi(S_k)\right) \right] \tag{34}$$

Unfortunately, the very large (i.e. high dimension) state space for the ODACS operational problem makes Eqn. (34) analytically intractable due to the curse of dimensionality (Powell, 2011). Researchers typically approximate the problem to obtain solutions. The next section presents an approximation of the objective function that coincides with an optimization-based joint AV-user assignment and empty AV repositioning operational policy.

8.4 Joint Assignment-Repositioning Operational Policy

This section describes the joint AV-user assignment and empty AV repositioning operational policy. Let V_k^I, V_k^P, V_k^U, V_k^I be the set of idle, pickup, in-use, and repositioning AVs at epoch k, respectively; where $V_k^I = \{V_j | \varphi_k^{V_j} = 0\}$, $V_k^P = \{V_j | \varphi_k^{V_j} = 1\}$, etc. This chapter presents two different joint assignment-repositioning operational policies. The first policy only allows idle AVs (V_k^I) and repositioning AVs (V_k^R) to have their plans adjusted (i.e. assigned to open user requests) at epoch k; whereas, the second policy also allows in-use AVs (V_k^U) to have their plans adjusted. Hence, let $V_k' = \{V_k^I, V_k^R\}$ and $V_k' = \{V_k^I, V_k^R, V_k^U\}$ be the controllable AVs in the former and later policies, respectively.

The joint assignment-repositioning operational policies involve solving the decision problem in Eqn. (35)-(37) and Eqn. (30)-(33), every decision epoch k, where t_{ij}^t and d_{ij} are the current travel time and distance between V_j and C_i, respectively. For notational simplicity, the variables do not include the decision epoch index.

$$\min_{x_{ij},y_{jh}} \sum_{i\in C}\sum_{j\in V'} x_{ij}\{c^{VOT}(t_{ij}^t - w_i) + c^{EDCR}(d_{ij}) - r^{asgn}\} + c^{busy}\sum_{i\in C}\sum_{j\in V^U} x_{ij} \\ + c^I \sum_{h\in Z} z_h + c^{EDCR}\sum_{j\in V'}\sum_{h\in Z} y_{jh}d_{jh} \tag{35}$$

$$I^h - \sum_{j\in V} y_{jh} - I^{min} \le z_h \qquad \forall h \in Z \tag{36}$$

$$z_h \ge 0 \qquad \forall h \in Z \tag{37}$$

Constraints in Eqn. (30)-(33)

The objective function contains several terms. The first large term includes the remaining wait time penalty associated with assigning V_j to C_i as well as the reward for assigning an AV to a user C_i with elapsed wait time w_i. It also includes the empty distance cost between V_j and C_i as well as the reward for any AV-user assignment. The second term penalizes the assignment of an in-use AV ($\varphi_k^{V_j} = 2$) to an open user request. This is done in order to account for the increased uncertainty of the in-use AVs availability to pick up the user compared to empty AVs. The third term penalizes the fleet for leaving an imbalance in subregion Z_h. Finally, the last term represents the empty distance cost between V_j and Z_h. The objective function implicitly makes trade-offs between assigning AVs to open requests now, reducing subregion imbalances now, and waiting until later (when other AVs become available) to assign AVs to open requests or balance subregions.

The parameters c^{VOT}, c^{EDCR}, r^{asgn}, c^{busy}, and c^{I} are the value of time, empty distance cost rate, reward for any AV-user assignment, penalty for assigning an *in-use* AV to an open request, and penalty for leaving an imbalance in a subregion, respectively. These parameters ensure all the terms in the objective function are in monetary units. The minimal imbalance parameter I^{min} allows a subregion to have a minimal imbalance before impacting the objective function.

The constraints in Eqn. (36) and Eqn. (37) ensure that z_k takes a value greater than or equal to the expected supply-demand imbalance, and zero, respectively. Fortunately, the constraint matrix (Eqn. (30)-(33) and Eqn. (36)-(37)) is totally unimodular; therefore, the linear relaxation of the integer program always produces integer solutions. Hence, even for large instances of this problem, solutions can be obtained in a reasonable amount of time. This is quite beneficial as the fleet needs to repeatedly resolve the problem every I^{d}.

8.5 Experimental Design

This section presents the experiments designed to compare the joint assignment-repositioning operational policy to other operational policies.

User Request Data

To obtain a spatio-temporal distribution for user requests, this study treats the taxi trips in the NYC taxi data (NYC Taxi & Limousine Commission, 2017) as user requests for an ODACS.

Ideally, this study would employ free-floating carshare data, as taxi trips are certainly different than carshare trips; however, carshare data was not accessible. The study uses data from the first fifteen days of April 2016 and includes a 10% sample of user requests for these days. The NYC taxi data includes the longitude and latitude of taxi trip origins and destinations. The taxi-user pickup time is treated as the user's request time in this chapter. The requested usage time for each user request is determined as follows. First, users are randomly assigned to a base usage time in which users are equally likely to be assigned to any usage time between 15-min and 3.0-hr in increments of 15-min. Second to determine a user's actual usage time, the algorithm draws from a normal distribution with mean μ_u and σ_u and subtracts this value from the base usage time.

Parameter Settings and Scenarios

Table 8-1 displays the parameter settings in the computational analysis that do not vary across scenarios. The AVs travel at speed $v = 5m/s$ since this is the average taxi speed in Manhattan. The simulations run from 2:30am to 12:30am (i.e. $t_{max} = 10\ hr$); whereas user requests only enter the system between 3:00am and 10:00am allowing the fleet to serve all users. This time frame incorporates the morning peak while allowing the AVs to reposition themselves before the morning peak to serve future user requests. The inter-decision time I^d is 30 seconds, a reasonable time for an on-demand service. The prediction horizon h^p is 30 minutes.

The value of time c^{VOT} and the empty distance cost rate c^{EDCR} used in the objective function and system cost evaluation metrics are \$23/hr. and \$0.31/km, respectively. These values coincide with estimates of value of time in the literature and the U.S. governmental mileage rate. The reward for assigning an AV to open user request r^{asgn} is \$5.0/user. Given a value of time of \$0.31/km, AVs will not be assigned to new user requests if the AVs are more than 16.1 km away from the new request. A smaller assignment reward value would increase user wait times but decrease empty fleet kilometers.

The imbalance penalty c^I is \$3.0/user, indicating empty AVs will only be considered for repositioning to an imbalanced subregion, if the AVs and subregions are less than 9.7 km away from each other. This chapter sets the AV-user assignment reward to a higher value than the imbalance penalty to prioritize serving open user requests that are known over probabilistic future requests. The minimum imbalance I^{min} is 1 user per subregion to provide some a buffer before repositioning AVs to subregions.

The model also adds a penalty of \$0.15/assignment for assigning in-use AVs to open user requests. This parameter incentivizes the assignment of idle and repositioning AVs over in-use AVs. The difference between requested usage time and actual usage time is distributed normally with mean μ_u 10 minutes and standard deviation σ_u 3 minutes.

Table 8-1: Parameter Settings that do not Vary across Scenarios

Parameter	Math Notation	Value	Units
Vehicle Speed	v	5	meters/sec
Simulation Length	t_{max}	10	hr.
Inter-decision Interval	I^d	30	sec
Prediction Horizon	h^p	30	min
Value of Time	c^{VOT}	23	\$/hr.
Empty Distance Cost Rate	c^{EDCR}	0.31	\$/km
Assignment Reward	r^{asgn}	5.0	\$/user
Imbalance Penalty	c^I	3.0	\$/user
Minimum Imbalance	I^{min}	1	user
In-use Penalty	c^{busy}	0.15	\$/assignment
Mean Usage Time Difference	μ_u	10	min
Std. Dev. Usage Time Difference	σ_u	3	min

The values in Table 8-1 are fixed across all scenarios/simulations in the computational analysis. However, the experimental design varies fleet size across simulation experiments. The three fleet sizes $|\mathcal{V}|$ are 1750, 2000, and 2250. Additionally, the experimental design tests all operational strategies using fifteen separate days of NYC taxi data.

Other Operational Strategies

The analysis in this chapter aims to compare the optimization-based joint assignment-repositioning operational strategy with other operational strategies. This chapter treats the operational strategy of assigning new user requests immediately to the nearest idle AV as the baseline strategy. Two other operational strategies involve solving an AV-user assignment decision problem. In one case, only idle AVs are considered; in the other strategy, idle and in-use AVs are considered. These strategies are similar to the ones in Chapter 6 for the *on-demand SAMS without shared rides.*

8.6 Computational Results

This section presents computational results comparing the operational policies for the ODACS operation problem in terms of three metrics, namely, average user wait time, total empty fleet kilometers, and system cost. The system cost is a weighted combination of the first two metrics. The operational strategy 'Asgn NN' is a myopic policy that immediately assigns new user requests to the nearest idle AV. The strategies 'Asgn Idle' and 'Asgn IdleDrop' involve solving an AV-user assignment optimization problem every I^d**,** where 'Asgn IdleDrop' considers *in-use* AVs in the decision problem in addition to *idle* and *repositioning* AVs. Finally, the 'Asgn-Rep Idle' and 'Asgn-Rep IdleDrop' operational strategies represent the optimization-based joint assignment-repositioning operational strategy. The ‘Asgn-Rep IdleDrop’ considers *in-use* AVs in addition to *idle* and *repositioning* AVs.

Figure 8-1 displays the computational results. Each point in each graph represents the average over 15 scenarios (i.e. the first fifteen days of April 2016). Given that the standard deviation across days is relatively small, fifteen days (i.e. replications) was enough to reduce the standard error such that there were statistically significant differences between the strategies.

The top graph in Figure 8-1 shows that the joint assignment-repositioning operational strategies significantly outperform the myopic nearest neighbor strategy and the two optimization-based assignment-only strategies. Average user wait time is around 7-min, 4-min, and 3-min for fleet sizes of 1750, 2000, and 2500, respectively for the joint assignment-repositioning strategies; whereas, for the same fleet sizes, average user wait time is around 10-min, 7-min, and 6-min for the optimization-based assignment-only strategies.

The middle graph in Figure 8-1 shows that the joint assignment-repositioning strategies produce slightly more empty kilometers than the two optimization-based assignment-only strategies. This is not surprising as repositioning is likely to increase empty fleet kilometers in nearly all cases. The assignment-only strategies with idle, repositioning, *and* in-use AVs produces between 8,800 km and 9,300 km; whereas, the joint assignment-repositioning strategies generate more than 10,000 km of empty kilometers. This indicates an inherent trade-off in terms of operational costs and service quality facing fleet controllers when selecting operational strategies for the ODACS operational problem. However, the fleet controller can adjust the parameters in the joint assignment-repositioning objection function to reduce empty fleet kilometers.

The bottom graph in Figure 8-1 shows the joint assignment-repositioning strategies significantly outperform the assignment-only strategies across all fleet sizes in terms of overall system costs. In fact, the system costs for assignment-only strategies are 40-60% higher than the joint assignment-repositioning system costs. This suggest that the joint assignment-repositioning operational strategies provide significant advantages over even the most advanced assignment only strategies. The results seem to indicate only marginal differences between the two optimization-based joint assignment-repositioning strategies, suggesting that either including in-use AVs provides minimal benefit or the operational strategy needs to more effectively incorporate in-use AVs considering the uncertainty associated with their remaining usage time.

The objective function for the joint assignment-repositioning operational strategy provides significant flexibility in terms of adjusting to the preferences of the fleet controller. If fuel costs increase, and it is important to reduce empty kilometers, the fleet controller can adjust the parameter values to reflect this change in costs. Similarly, if the ODACS provider wants to focus on reducing user wait times, even at the expense of more empty fleet kilometers, the objective function can reflect this strategic change.

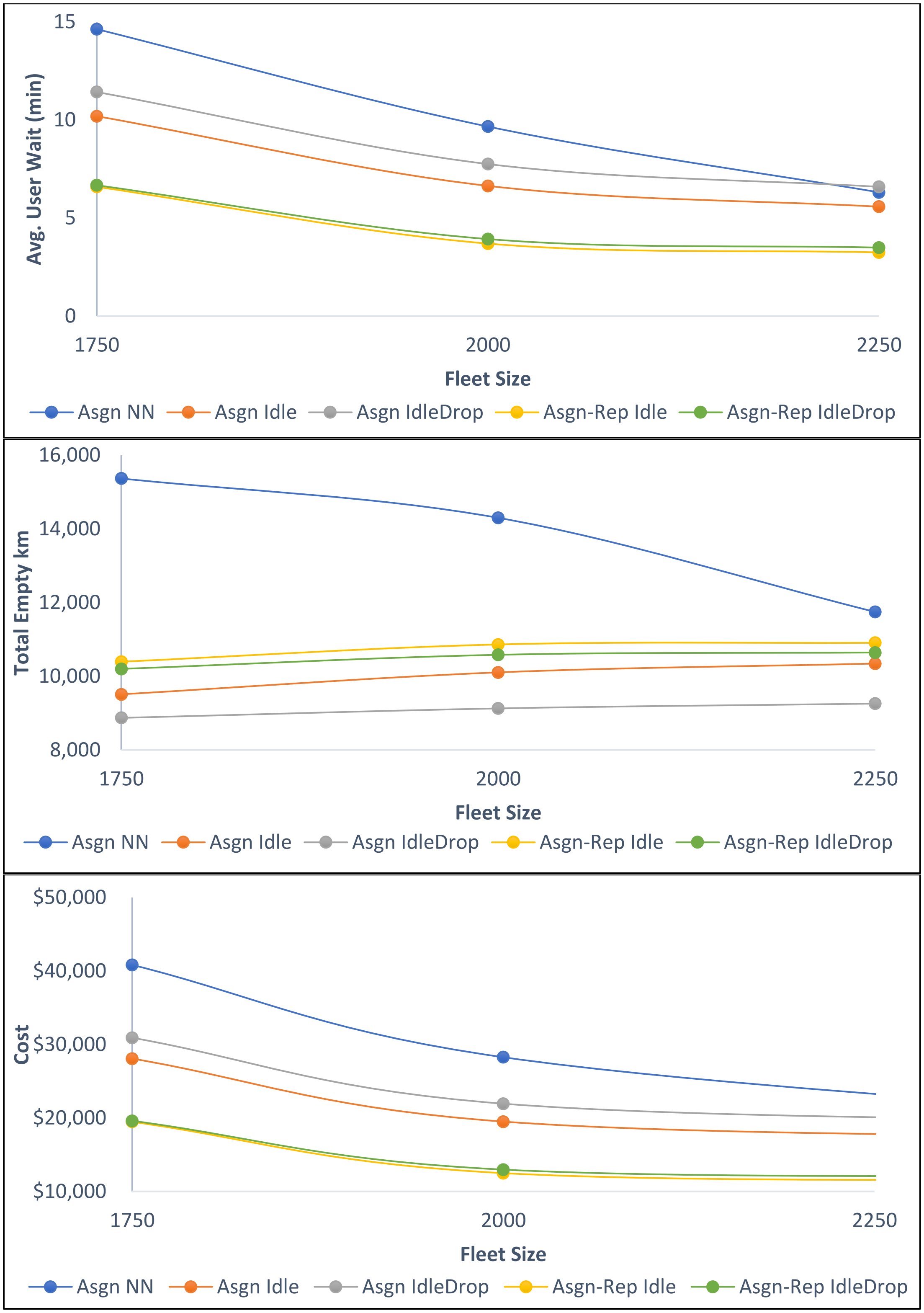


Figure 8-1: Comparison of operational strategies (line color) in terms of average user wait time (top), total empty fleet kilometers (middle) and system cost (bottom) across fleet sizes (x-axis)

8.7 Conclusion

This chapter examines the *on-demand autonomous carsharing service (ODACS)* operational problem, which is a stochastic dynamic control problem (SDCP). This chapter presents a modeling framework for the ODACS operational problem that captures the evolution of the system state through time and sequential nature of decisions. At each decision epoch, the available AVs (i.e. *idle* AVs, *repositioning* AVs, and in some cases *in-use* AVs) can either be (i) assigned to open user requests, (2) repositioned to a subregion centroid, (3) told to remain in their current location. The problem's objective is to minimize a weighted combination of average user wait times and empty fleet kilometers.

To address this SDCP, this chapter introduces a joint AV-user assignment and empty AV repositioning strategy that involves solving a decision problem at each decision epoch. The decision problem via a joint assignment-repositioning objective function implicitly trades-off the benefits of immediately serving user requests, with the downstream benefits of repositioning empty AVs to decrease supply-demand imbalances and subsequently serve future user requests efficiently. The chapter compares this operational strategy to optimization-based AV-user assignment strategies that do not incorporate repositioning. The computational results show sizable performance advantages for the joint assignment-repositioning strategy over the assignment-only strategy, in terms of user wait times and systems costs, the latter metric being a weighted combination of user wait times and empty fleet kilometers.

An ODACS combines the benefits of existing MOD services and existing non-autonomous services, including on-demand service where AVs pick up users at their origin and drop them off at their destinations, while still allowing them to reserve an AV for a user-specified time-slot to make multiple trips and store items between trips. The benefits of this service suggest that it may be an attractive SAMS in the future. Hence, from a mobility service provider perspective, and a transportation system efficiency perspective, research on ODACSs is critical to their future success.

Chapter 9 Impact of Spatio-Temporal Demand Forecast Aggregation on the Operational Performance of Shared Autonomous Mobility Fleets[8]

9.1 Overview

Fleet operators rely on forecasts of future user requests to reposition empty vehicles and efficiently operate their vehicle fleets. In the context of an on-demand shared-use autonomous vehicle (AV) mobility service (SAMS), this study analyzes the trade-off that arises when selecting a spatio-temporal demand forecast aggregation level to support the operation of a SAMS fleet. In general, when short-term forecasts of user requests are intended for a finer space-time discretization, they tend to decrease in quality. However, holding forecast quality constant, more disaggregate forecasts provide more valuable information to fleet operators.

To explore this trade-off, this study presents a flexible methodological framework to evaluate and quantify the impact of spatio-temporal demand forecast aggregation on the operational efficiency of a SAMS fleet. At the core of the methodological framework is an agent-based simulation that requires a demand forecasting method and a SAMS fleet operational strategy. This study employs an offline demand forecasting method, and an online joint AV-user assignment and empty AV repositioning strategy. Using this forecasting method and fleet operational strategy, as

[8] An article is under 2nd review by Dandl, Hyland, Bogenberger, and Mahmassani that parallels this chapter.

well as Manhattan, NY taxi data, this study simulates the operations of a SAMS fleet across various spatio-temporal aggregation levels.

Results indicate that as demand forecasts (and subregions) become more spatially disaggregate, fleet performance improves, in terms of user wait time and empty fleet miles. This finding comes despite demand forecast quality decreasing as subregions become more spatially disaggregate. Additionally, results indicate the SAMS fleet significantly benefits from higher quality demand forecasts, especially at more disaggregate levels.

9.2 Motivation

The growth of shared-use mobility services and the availability of large data sources (e.g. taxi and carsharing data) has prompted significant research in the transportation literature. The advent of AVs and their expected inclusion in mobility service fleets has further motivated research relating to the operation and management of SAMSs. This research falls in two main areas: forecasting demand/user requests (i.e. modeling arrival processes) and developing operational policies/strategies to efficiently operate a SAMS fleet dynamically.

The existing literature largely treats these two problems independently. To address the forecasting problem, researchers are developing and comparing demand forecasting methods (Sayarshad and Chow, 2016). To address the problem of operating SAMS fleets efficiently, researchers are developing strategies to assign AVs to user requests (Alonso-Mora et al., 2017; Hyland and Mahmassani, 2018; Maciejewski et al., 2016) and reposition empty AVs (Dandl and Bogenberger, 2018b; Fagnant and Kockelman, 2014; Hörl et al., 2017; Pavone et al., 2012; Sayarshad and Chow, 2017; Spieser et al., 2016). The repositioning strategies rely on forecasts of future demand to reposition empty AVs; hence these two SAMS subproblems are inherently interconnected.

Figure 9-1 shows the relationship between demand forecasting (i.e. predictive analytics) and SAMS operational decision-making (i.e. prescriptive analytics). Predictive analytics methods convert ‘raw’ data into (demand) forecasts; whereas, prescriptive analytics (i.e. optimization) methods rely on these demand forecasts to prescribe informed (operational) decisions (IBM, 2017). Hence, the efficient operation of a SAMS fleet requires reliable demand forecasts. Moreover, forecasts only provide real value to a SAMS provider if they improve decision making and fleet performance.

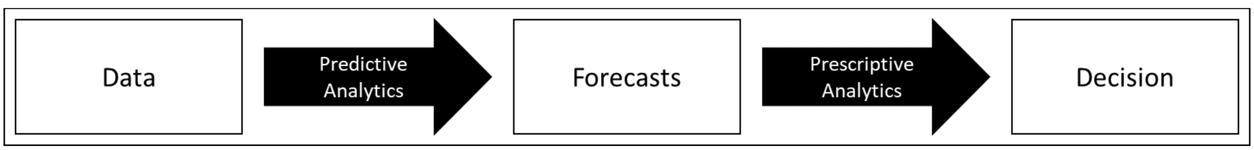


Figure 9-1: Schematic of process to convert data into better decisions

Motivated by the inherent interconnection between demand forecasts and the operational performance of SAMS fleets, this chapter aims to connect these two research areas. Mobility service providers need to consider both problems (jointly). Specifically, this chapter aims to evaluate and quantify the impact of spatio-temporal demand forecast aggregation on the operational performance of a SAMS fleet. Given the sizeable market share of existing on-demand mobility services that do not allow shared rides (e.g. UberX, traditional Lyft, and taxi services), this chapter analyzes an *on-demand SAMS without shared rides*, defined in Chapter 6.

9.3 Background

Research illustrates that the utilization of advanced (deterministic or stochastic) information can improve the operation of mobility services. In the context of goods transport, Yang et al. (2004)

present the generic real-time truckload pickup and delivery problem and present computational results as a function of advanced information about demand requests. Tjokroamidjojo et al. (2006) and Jaillet and Wagner (2006) quantify the value of advanced deterministic information (i.e. known future requests) in dynamic freight routing problems. The on-demand SAMS modeled in this chapter does not allow users to request rides in advance; therefore, the SAMS fleet cannot obtain advanced (deterministic) information. However, predictive analytics methods and big data can help SAMS operators forecast demand and reposition vehicles based on demand predictions, thereby reducing user wait times.

In the context of carsharing, Weikl and Bogenberger (2015) introduce an algorithm to relocate vehicles, based on forecasts of future demand, in order to maximize profit. In goods transport, Ichoua et al. (2006) use demand forecasts to decide whether a vehicle should wait in its current position for a future demand before continuing its planned tour. Some SAMS studies introduce empty vehicle repositioning strategies (Dandl and Bogenberger, 2018b; Fagnant and Kockelman, 2014; Hörl et al., 2017; Pavone et al., 2012; Sayarshad and Chow, 2017; Spieser et al., 2016); however, these studies do not focus on the implications of demand forecast aggregation and/or quality on fleet performance.

The existing literature includes short-term demand forecasting studies related to carsharing (Müller and Bogenberger, 2015), taxi (Ihler et al., 2006; Moreira-Matias et al., 2013), and public transportation (Zhong et al., 2016). In their survey and comparative analysis of taxi user arrival process models, Sayarshad and Chow (2016) categorize forecast methods into offline models and online models. Offline models rely entirely on historic data; whereas, online models utilize real-time data. Sayarshad and Chow (2016) evaluate the prediction quality of two offline and three online forecast models using New York taxicab data. Recent demand forecasting research incorporates new features from other data sources (e.g. social media) to further improve the quality of online models (Chaniotakis et al., 2016; Tong et al., 2017). In a more general analysis (i.e. broader than transportation), Zotteri et al. (2005) present an in-depth analysis of the impact of aggregation level on forecasting performance.

9.4 Research Problem and Hypothesis

The purpose of this study is to evaluate and quantify the impact of spatio-temporal demand forecast aggregation on the operational performance of SAMS fleets. There is an inherent trade-off in the selection of a spatio-temporal aggregation level. From an operational standpoint, holding forecast reliability constant, more disaggregate spatio-temporal forecasts – for smaller subregions – provide the fleet more valuable information. For example, knowing three users will request rides in a 100 m^2 area between 9:00am and 9:05am is more valuable than knowing three users will request rides in a 1000 m^2 area between 9:00am and 9:30am. However, it is likely that short-term SAMS demand forecast errors will increase as forecasts become very disaggregated in space and time due to the law of large numbers (statistical variability increases as the number of items to forecast decreases) and the underlying demand generation process (Makridakis, 1988). Given the inherent trade-off associated with choosing a spatio-temporal aggregation level, this study's working hypothesis is that:

- SAMS fleet performance will initially increase as forecasts (and subregions) become more disaggregate; however, eventually, SAMS fleet performance will decrease, or at least stagnate, as forecasts and subregions become progressively more disaggregate.

This chapter also aims to determine the optimal spatio-temporal demand forecast aggregation level to most efficiently operate a SAMS fleet. However, the optimal spatio-temporal aggregation level depends on a multitude of factors including the demand forecasting method, the SAMS operational strategy, and even the characteristics (e.g. density of user requests) of the service area. Hence, this chapter introduces a flexible methodological framework that other researchers and mobility service providers can employ to determine the optimal spatio-temporal aggregation level for their own forecasting method, fleet operational strategy, and service area.

To obtain a quasi-upper bound on the operational performance of the SAMS fleet, the chapter runs experiments wherein the fleet has perfect demand forecasts, across all spatio-temporal aggregation levels. To obtain a quasi-lower bound, the chapter runs experiments where the fleet has no information about future demand forecasts.

9.5 Research Methodology

To perform the computational analysis and test the hypothesis, this study employs an agent-based simulation tool. The simulation tool models the operations of an AV fleet, employs an algorithm to efficiently assign AVs to open user requests, and uses demand forecasts to proactively reposition AVs to serve future user requests. After providing an overview of the simulation framework, this section details the user requests, the demand forecast model, and the SAMS fleet operational strategy employed in this study. The agent-based simulation model in this study follows the general three-component framework for modeling SAMSs that includes a demand (i.e. traveler request) generator, an SAMS fleet operator/dispatcher, and some representation of the transportation network (Levin et al., 2017b; Rigole, 2014).

Agent-based Simulation Framework

Figure 9-2 displays a flowchart of the agent-based simulation tool. The simulation tool is time-driven and updates the position and status of AVs and users every time step. To initialize the simulation, the current time, τ, is set to zero, the AVs are positioned throughout the service region, and the statuses of all AVs are set to *idle*.

The simulation first updates the current time ($\tau \leftarrow \tau + \Delta\tau$) by the simulation time step ($\Delta\tau$) and then checks if τ is less than the length of the simulation period T. If $\tau \geq T$, the simulation ends, otherwise the simulation moves en-route pickup AVs, en-route drop-off AVs, and en-route repositioning AVs one step ($\Delta\tau \times v$, where v is vehicle speed) closer to their assigned destination.

After the simulation moves the vehicles, it checks for new user requests with a request time $r_i = \tau$, where r_i is the request time of user i. Then the simulation checks to see if it is time to assign AVs to open user requests and reposition AVs to different subregions. Every I^d, the inter-decision time interval length, the fleet simultaneously assigns and repositions AVs. Figure 9-2 shows that the joint assignment-repositioning strategy requires spatio-temporal demand forecasts, which are a key input in the operational strategy, as they determine how many, when, and where AVs should reposition.

The fleet only assigns and repositions AVs every I^d for strategic reasons and computational constraints. Strategically, it is often beneficial to allow user requests to queue before assigning AVs to them, especially, if AV-user assignments are final (i.e. if AV diversions and user reassignments are not allowed). The constraint comes from the fact that it can take more than a few seconds to solve a decision problem that involves assigning and repositioning large numbers of AVs.

After moving the AVs, checking for new user requests, and assigning AVs to user requests and subregions, the simulation updates the system state via changing the statuses of AVs, users, and subregions, if necessary. For example, if an AV reaches its drop-off point, the simulation changes the status of the AV from *en-route drop-off* to *idle*.

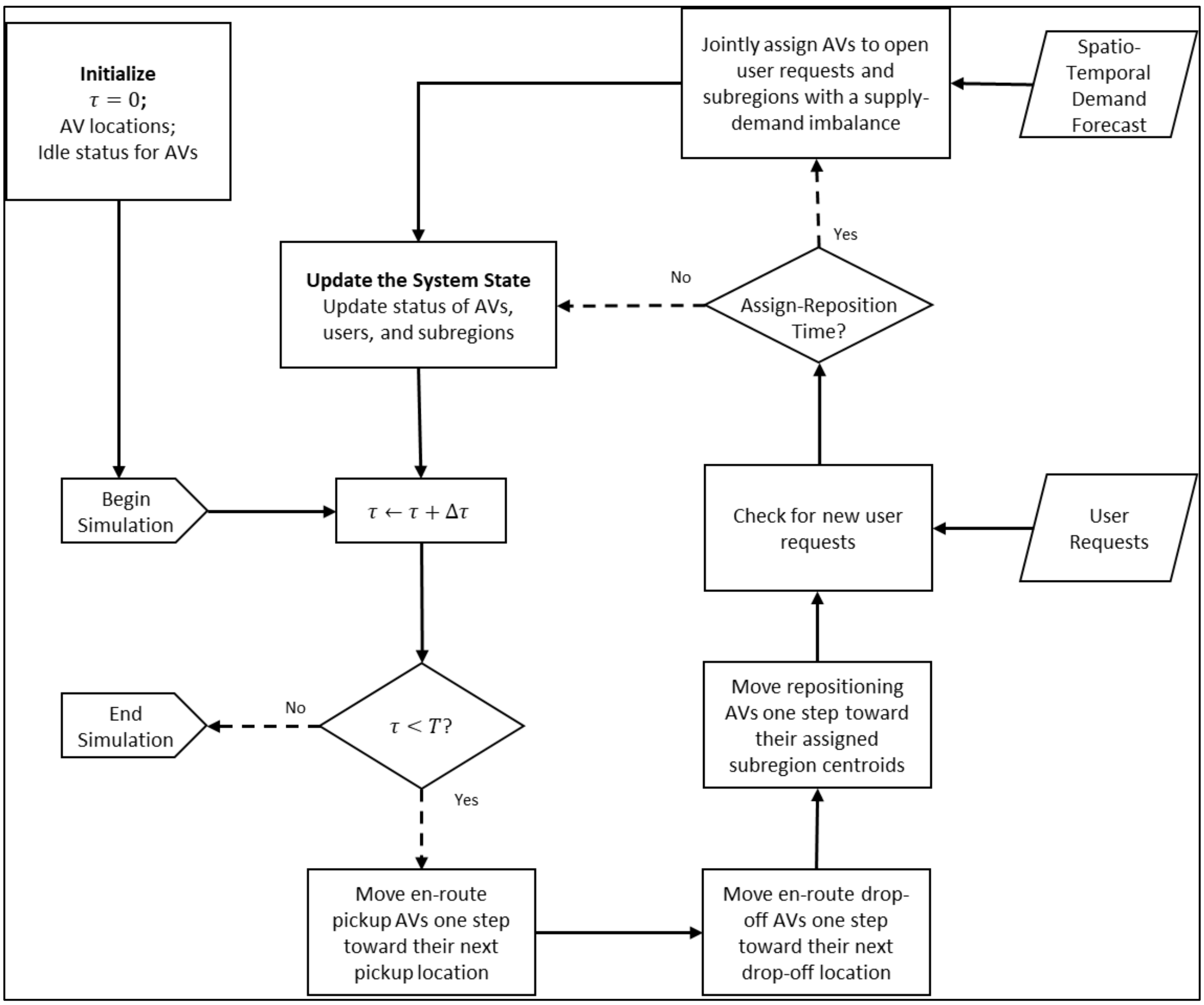


Figure 9-2: Simulation framework

The simulation ends when $\tau = T$, even if AVs are still active and users are still unserved. The simulation can output metrics for individual AVs, users, and subregions, such as wait time and (empty) vehicle miles. As this chapter focuses on the performance of the SAMS fleet across different spatio-temporal aggregation levels, the computational analysis section presents performance statistics at the system level, such as average user wait time and empty fleet miles.

User Requests

The main input to the agent-based simulation model is the set of user requests. Each user request i includes an origin (o_i), destination (d_i), and request time (r_i). In the simulation, the fleet becomes aware of each user and her origin and destination, at her request time (i.e. when the user requests a ride on her smartphone). This chapter assumes that the SAMS serves every user request within the service region; moreover, it assumes users will wait indefinitely to be served.

Demand Forecasts

As described previously, fleet repositioning algorithms require forecasts of future demands. This chapter employs two sets of demand forecasts to analyze the impact of spatio-temporal aggregation on SAMS fleet performance. The first set of forecasts come from a simple time-varying Poisson forecast method based on historical demand (Ihler et al., 2006; Moreira-Matias et al., 2013; Sayarshad and Chow, 2016; Tong et al., 2017), whereas, the second set of forecasts are *perfect* demand forecasts.

Demand Forecasting Model

Like Moreira-Matias et al. (2013), this study uses a simplified version of the time-varying Poisson model in Ihler et al. (2006), which exploits weekly periodicity in demand to make forecasts for future days. The model in this chapter does not include seasonality terms because it only uses three months of data. Historical request data $\{o_i, d_i, r_i\}_{\forall i \in C}$ are aggregated into spatio-temporal bins. The underlying assumption is that (for example) the requests on Sunday between 5:00pm and 5:30pm will be similar to the historic average of requests on previous Sundays between 5:00pm and 5:30pm. Hence, the forecasted trip origin count for subregion k during period h on day-of-the-week d is based on the historical average of trip counts in subregion k during period h on day-of-the-week d. Although more advanced methods tailored to specific problem instances can likely produce better results, this study employs the 'historical average' model or time-varying Poisson model because of its wide-use in practice and in the literature due to its ease of implementation.

The Poisson distribution is defined below, where λ is the rate of new user requests entering the system, n is the number of new user requests, and $P(n;\lambda)$ is the probability of exactly n new user requests entering the system over a specified time period, given rate λ.

$$P(n;\lambda) = \frac{e^{-\lambda}\lambda^{n}}{n!}$$

However, the rate λ is not time-invariant and space-invariant in real-world shared-use mobility services; rather, it varies across space and time. Similar to the model in Moreira-Matias et al. (2013), this study assumes the time- and space-variant rate $\lambda_k(t)$ is a function of the day of the week $d(t)$, the period of the day $h(t)$, and the subregion k. This functional relationship is displayed below, where $\lambda_{k,0}$ is the average rate over the week in subregion k, $\delta_{k,d(t)}$ is the relative change for day-of-the-week $d(t)$ in subregion k, and $\eta_{k,d(t),h(t)}$ is the relative change for period $h(t)$ on day-of-the-week $d(t)$ in subregion k.

$$\lambda_k(t) = \lambda_{k,0} \times \delta_{k,d(t)} \times \eta_{k,d(t),h(t)}$$

This study varies the size of the period $h(t)$ and the size of each subregion k in order to determine the impact of temporal aggregation and spatial aggregation, respectively, on demand forecast quality and SAMS fleet performance. The parameters in the formula are calibrated using historical trip request data for various period $h(t)$ sizes and subregion k sizes. Multiplying $\lambda_k(t)$ by the size of the period $h(t)$ gives the expected number of new user requests in subregion k, on day $d(t)$, during period $h(t)$ that is used by the SAMS fleet operator.

Perfect Demand Forecasts

Obtaining perfect demand forecasts requires aggregating the actual request data into different spatio-temporal bins. If spatial aggregation is set at the census tract level and temporal aggregation is set at the one-hour level, then, with perfect forecasts, the SAMS fleet knows the exact number of users who will request service originating at each census tract every hour of the day. However, the SAMS fleet does not know the exact location within the census tract, nor does it know the exact time within the one-hour interval the requests will occur. Hence, more disaggregate subregions and time intervals provide the SAMS fleet more valuable information.

SAMS Fleet Operational Strategy

This section describes a SAMS fleet operational strategy that jointly assigns AVs to open user requests and repositions AVs between subregions.

Let V denote the set AVs in the SAMS fleet and let $j \in V$ denote an AV in the fleet. Moreover, let V^I, V^P, V^D, and V^R be the set of idle, en-route pickup, en-route drop-off, and en-route repositioning AVs respectively; $V = \{V^I, V^P, V^D, V^R\}$.

Similarly, let C denote the set of open user requests (meaning, they have not been assigned to an AV yet) and let $i \in C$ denote an open user request. If τ is the current time and r_i is the request time of user i, then user i's elapsed wait time (w_i) is $w_i = \tau - r_i$.

Additionally, let R denote the set of subregions in the service area, and let $k \in R$ denote a subregion. The expected imbalance between AVs and open user requests in subregion $k \in R$ over the prediction horizon h^p is denoted I_k. The expected imbalance I_k is determined by taking the difference between expected future demand and planned future supply in subregion k between the current time τ and the end of the prediction horizon $\tau + h^p$. The expected future demand is the sum of:

- the number of open user requests currently in subregion k; and
- the expected number of future requests in subregion k over the prediction horizon h^p (this value comes from the demand forecasts).

The planned future supply is the sum of:

- the number of repositioning AVs and idle AVs currently in subregion k;
- the number of en-route drop-off and en-route pickup AVs assigned to users who have destinations in subregion k (the AVs must drop off their users in subregion k within the prediction horizon h^p);

The current distance between AV j and open user request i is denoted d_{ij}. The distance between AV j and the demand-weighted centroid of subregion k is denoted d_{jk}.

To solve the stochastic dynamic problem of operating a SAMS fleet, this chapter employs a rolling-horizon solution approach, wherein every I^d (the inter-decision time) the fleet solves an optimization problem. In this chapter, the fleet can only control the AVs that are currently idle V^I

or repositioning V^R. In fact, from the perspective of the fleet, at the decision epoch (every I^d) there is no difference between AVs that are currently repositioning and AVs that are currently idle. Hence, let $V' = \{V^I, V^R\}$ denote the subset of AVs the algorithm can (i) assign to open user requests, (ii) reposition to subregions, or (iii) choose to be idle.

There are several key differences between the solution approach in this chapter and the solution approach for the *on-demand SAMS without shared rides* in Chapter 6. First, the solution approach in this chapter incorporates AV repositioning to subregion centroids throughout the service region, whereas, Chapter 6 does note. Second, the solution approach in Chapter 6 requires either all AVs to be assigned to a user request (if there are more open user requests than available AVs) or all user requests to be assigned to an AV (if there are more available AVs than open user requests). The solution approach in this chapter does not include this constraint pair, rather, the objective function includes a reward term for assigning AVs to user requests. Third, unlike Chapter 6, the solution approach in this section does not allow en-route pickup AVs to be diverted to other travelers nor does it allow assigned users to be reassigned to a new AV.

To model the decision problem mathematically, let x_{ij} equal one if AV j is assigned to pick up user i, and zero otherwise. Moreover, let r_{jk} equal one if AV j is assigned to reposition to subregion k, and zero otherwise. Equation (38) displays the objective function driving the fleet controller's decisions; whereas, Eqn. (39)-(42) constrain the decision set.

$$\min_{x_{ij}, r_{jk}} \begin{gathered} c^{ED} \sum_{i \in C} \sum_{j \in V'} x_{ij} d_{ij} - r^{asgn} \sum_{i \in C} \sum_{j \in V'} x_{ij} + c^I \max\left(0, \sum_{k \in R} \left(I_k - \sum_{j \in V'} r_{jk}\right) - I^{min}\right. \\ + c^{ED} \sum_{j \in V'} \sum_{k \in R} r_{jk} d_{jk} - c^{VOT} \sum_{i \in C} w_i \sum_{j \in V'} x_{ij} \end{gathered} \tag{38}$$

s.t.

$$\sum_{i \in C} x_{ij} + \sum_{k \in R} r_{jk} \leq 1 \qquad \forall j \in V' \qquad (39)$$

$$\sum_{j \in V'} x_{ij} \leq 1 \qquad \forall i \in C \qquad (40)$$

$$x_{ij} \in \{0,1\} \qquad \forall i \in C, \forall j \in V' \qquad (41)$$

$$r_{jk} \in \{0,1\} \qquad \forall j \in V', k \in R \qquad (42)$$

The objective function contains five separate terms that are associated with a penalty or a reward. The first term is a penalty term that denotes the cumulative distance between each newly assigned AV j and the user i it will pick up. The second term rewards the fleet for assigning an AV j to an open user request i. The third term penalizes the fleet for allowing an imbalance, greater than the minimum imbalance parameter I^{min}, in subregion k. The fourth term is a cost term that denotes the cumulative distance between each AV j and the centroid of subregion k it is assigned. The fifth term further rewards the fleet for assigning AVs to user requests with a long elapsed wait time.

The parameters set $(c^{ED}, r^{asgn}, c^{I}, c^{VOT})$ convert units of empty vehicle distance, passengers assigned, expected subregion imbalances, and elapsed wait time into monetary units. The objective function implicitly makes trade-offs between assigning AVs to open requests now, reducing subregion imbalances now, and waiting until later (when other AVs will become available) to assign AVs to open requests or balance subregions.

The constraint in Eqn. (39) ensures that each AV j is assigned to at most one open user request or subregion k. The constraint in Eqn. (40) ensures that no more than one AV is assigned to a single open user request. The constraints in Eqn. (41)-(42) ensure the two sets of decision variables take on binary values.

The third term in the objective function with the max() term is nonlinear. Fortunately, it is easy to convert this term into a linear integer programming problem. The term z_k in Eqn. (43) replaces the max() term in Eqn. (38). The constraints in Eqn. (44) and Eqn. (45) ensure that z_k takes a value greater than or equal to the original value in the max() term of Eqn. (38), and zero, respectively. The constraints in Eqn. (39)-(42) remain.

$$\min_{x_{ij},r_{jk}} c^{ED} \sum_{i \in C} \sum_{j \in V'} x_{ij} d_{ij} - r^{asgn} \sum_{i \in C} \sum_{j \in V'} x_{ij} + c^{I} \sum_{k \in R} z_k + c^{ED} \sum_{j \in V'} \sum_{k \in R} r_{jk} d_{jk} - c^{VOT} \sum_{i \in C} w_i \sum_{j \in V'} x_{ij} \tag{43}$$

s.t.

$$I_k - \sum_{j \in V} r_{jk} - I^{min} \leq z_k \qquad \forall k \in R \tag{44}$$

$$z_k \geq 0 \qquad \forall k \in R \tag{45}$$

Constraints Eqn. (39)-(42)

Fortunately, the constraint matrix – Eqn. (39)-(42) and Eqn. (44)-(45) – is totally unimodular; therefore, the linear relaxation of the integer program always produces integer solutions. Hence, even for large instances of this problem, solutions can be obtained in a reasonable amount of time. This is quite beneficial as the fleet needs to repeatedly resolve the problem every I^d.

Fleet Strategy without Demand Forecasts

As mentioned previously, this chapter aims to create a quasi-lower-bound on fleet performance via testing scenarios that only allow myopic operational strategies that do not consider demand forecasts to replicate the case in which short term demand forecasts are not available. Using the variable definitions described above, Eqn. (46)-(49) define the myopic user assignment strategy without AV repositioning. This math program parallels the formulation in Eqn. (38)-(42), except it does not include the repositioning terms and constraints.

$$\min_{x_{ij}} c^{ED} \sum_{i\in C} \sum_{j\in V^I} x_{ij} d_{ij} - r^{asgn} \sum_{i\in C} \sum_{j\in V^I} x_{ij} - c^{VOT} \sum_{i\in C} w_i \sum_{j\in V^I} x_{ij} \tag{46}$$

s.t.

$$\sum_{i\in C} x_{ij} \leq 1 \qquad \forall j \in V^I \tag{47}$$

$$\sum_{j\in V^I} x_{ij} \leq 1 \qquad \forall i \in C \tag{48}$$

$$x_{ij} \in \{0,1\} \qquad \forall i \in C, \forall j \in V^I \tag{49}$$

9.6 Experimental Design

NYC Taxi Data

This chapter utilizes taxi data from New York City provided by the NYC Taxi and Limousine Commission (2017). The yellow taxi data was filtered for trips starting *and* ending in Manhattan since this simplifies the process of aggregating data into subregions of different sizes. The simulation treats the recorded taxi trip start times in the NYC taxi data as the users' request times.

Averaging over all days in April 2016, the number of trips per hour varies between 2,500 trips per hour between 4:00am and 5:00am and more than 21,300 trips per hour between 06:00pm and 08:00pm. For trips per day, the mean is 314,796 trips and the standard deviation is 69,122 trips. The mean taxi trip length is 2.8 km (1.7 mi) with a standard deviation of 2.0 km (1.2 mi).

The origins and destinations of all taxi records were transformed into a metric system and create a minimum bounding rectangle. To create the largest forecast subregions, the short edge of the minimum bounding rectangle was cut in two and the long edge in eight pieces to create approximately square areas. To generate more disaggregate subregions, the edges of each subregion are cut in half. Figure 9-3 displays the resulting forecast subregions for four spatial aggregation levels. This method produces large differences in the number of trips per zone (i.e. a high coefficient of variation for daily trips per subregion) but provides an efficient means to test different spatial aggregation levels.

Since the simulation framework only allows movements along the x-axis and y-axis, the coordinate system for Manhattan, along with user's origins and destinations, were rotated to align with the gridded street network.

Simulating a full day and using a realistic spatio-temporal demand distribution adds practical value to the results presented in the next section. Moreover, given the natural spatio-temporal fluctuations in demand throughout a typical day, the SAMS fleet relies on demand forecasts to reposition AVs in advance of future demand surges. The demand forecast model was calibrated based on three months of historical data.

A preliminary analysis of the NYC taxi trip request data supports the choice of a historical average forecast model that segments the data by day-of-the-week. In the case of hourly demand forecasts for the entire borough of Manhattan, not segmenting by day-of-the-week results in a coefficient of variation (CV) for taxi trip count of 20%. The CV is between 3% (Wednesdays) and 7% (Saturdays) when the data are segmented by day-of-the-week.

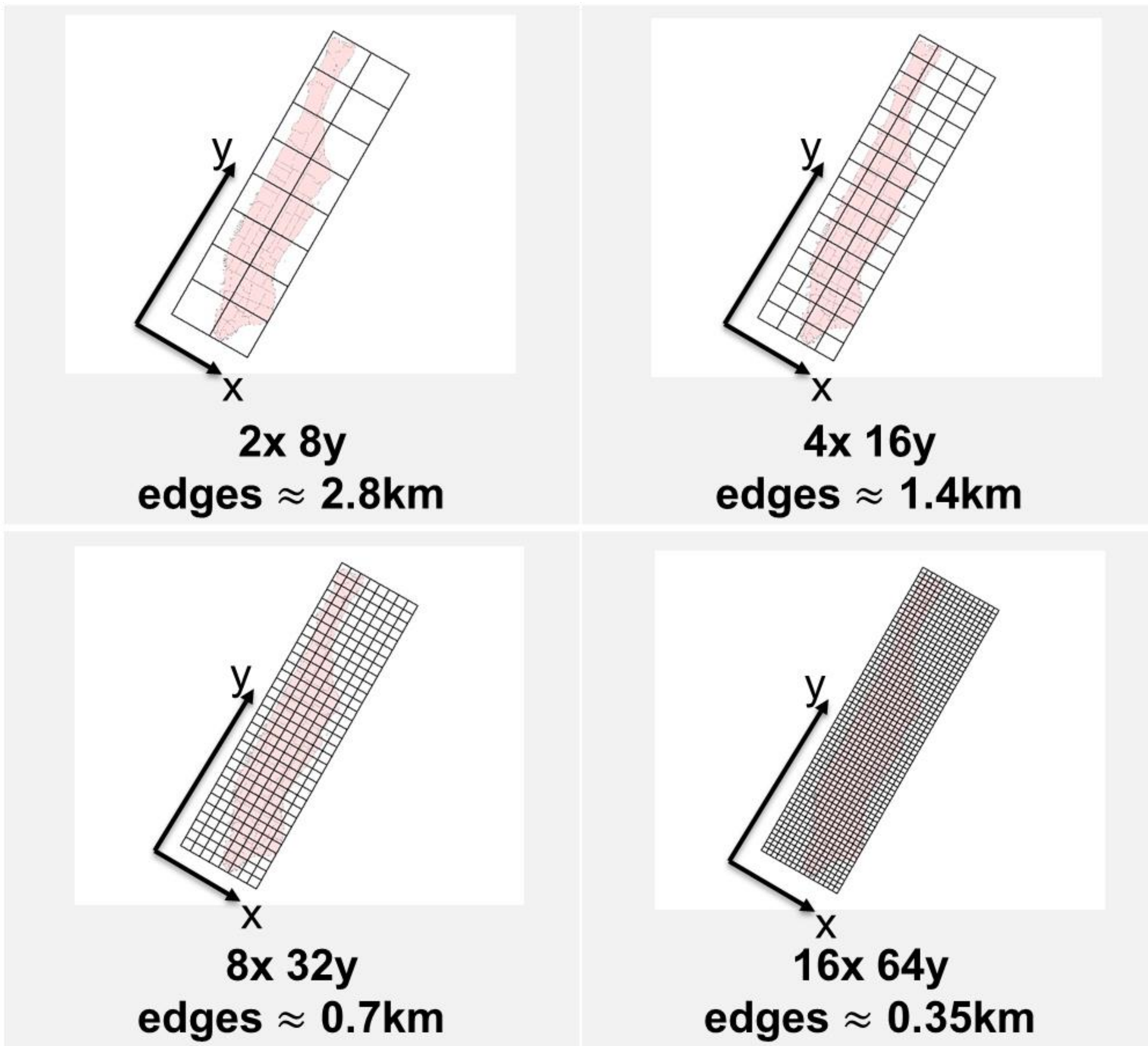
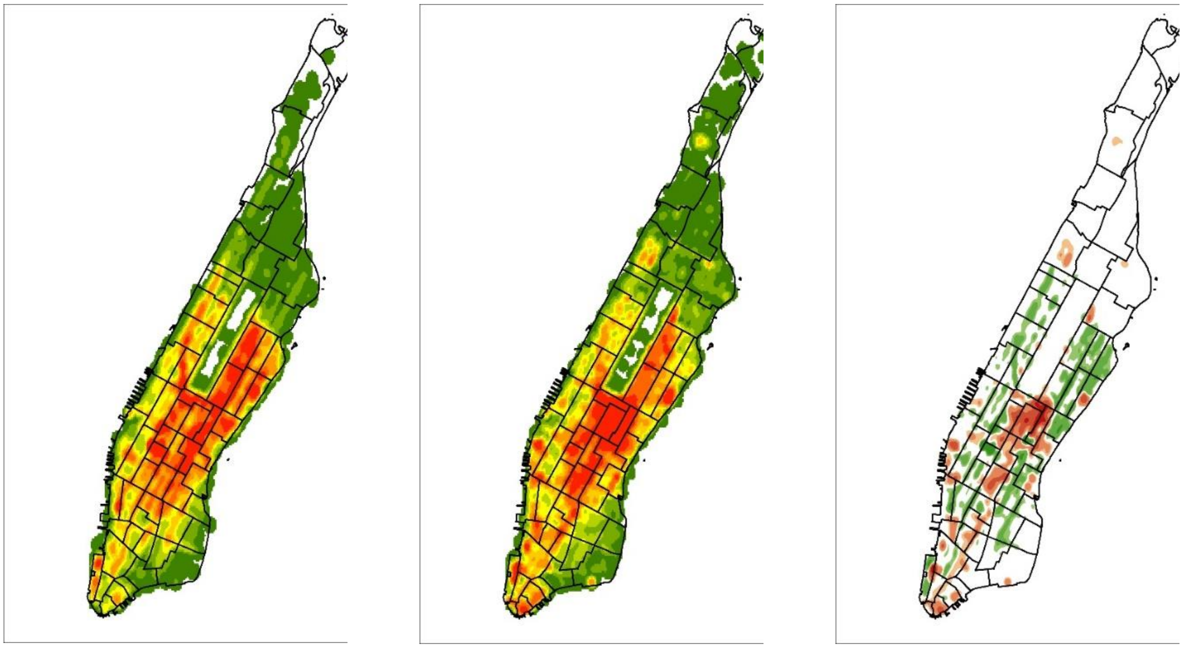


Figure 9-3: Subregion layout for different spatial aggregation levels (official NYC taxi subregions for Manhattan are drawn in the background)

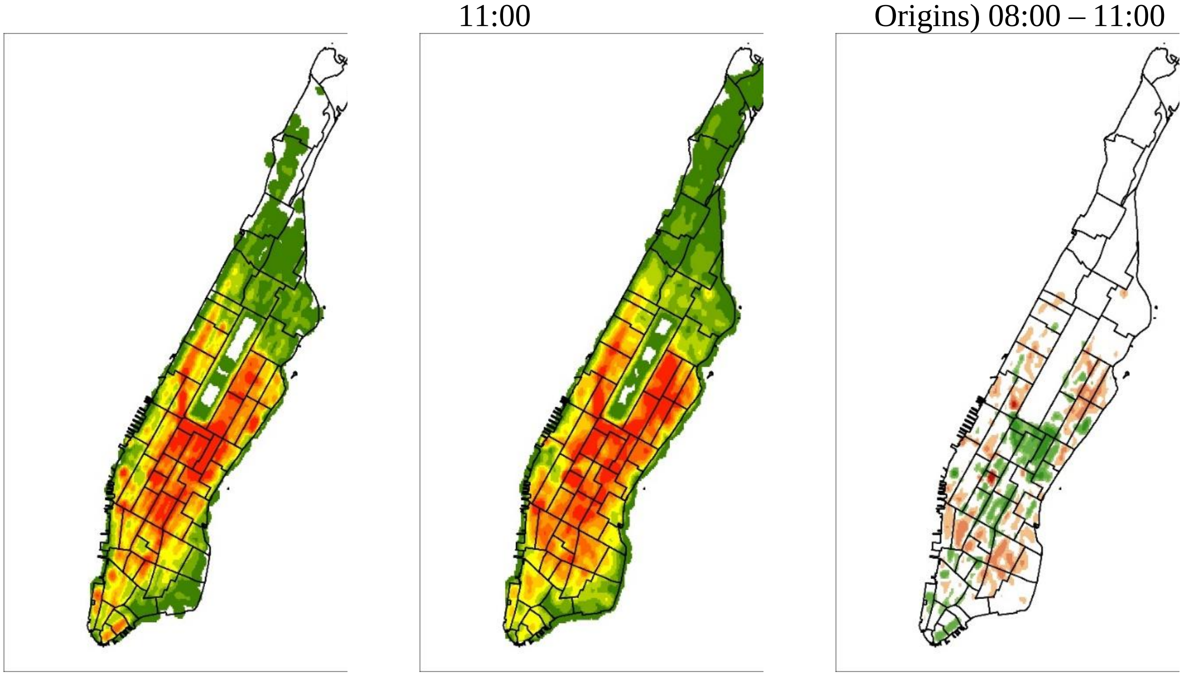
Figure 9-4 displays demand density across Manhattan during the morning (8:00am-11:00am) and evening (5:00pm-8:00pm) for taxi users. These plots were created using the ArcGIS 'kernel plot' function. The density plots on the left-side and in the middle display the density of trip origins, and destinations, respectively. The density plots on the right-side display the net difference between trip destinations and trip origins. In the density plots on the right-side, areas in red (green) denote areas where there are more (fewer) trips terminating than originating. As the fleet serves demand throughout the day, without repositioning AVs, red (green) areas are likely to have a surplus (deficit) of AVs. Hence, repositioning empty AVs, from surplus areas to deficit areas, should improve the operational performance of SAMS fleets.



a) Origins 08:00 – 11:00

b) Destinations 08:00 – 11:00

c) Net (Destinations – Origins) 08:00 – 11:00



d) Origins 17:00 – 20:00

e) Destinations 17:00 – 20:00

f) Net (Destinations – Origins) 17:00 – 20:00

Figure 9-4: Taxi trip density during the morning (a-c) and evening (d-f) on Wednesday, 2016-04-06 for trip origins (a and d), trip destinations (b and e), and net trips (c and f) wherein red (green) areas represent more (fewer) trip origins, trip destinations, and net trips, respectively

Parameter Settings

Table 9-1 displays the parameter settings in the computational analysis that do not vary across scenarios. Test simulations with 3,500 to 6,000 AVs indicate fleet size is a crucial parameter for this study. On the one hand, very small fleet sizes essentially preclude repositioning trips because all vehicles are continuously busy serving a growing queue of open user requests. On the other hand, very large fleet sizes allow the fleet to easily serve all user requests without any subregions ever experiencing a deficit of AVs. Test simulations indicated that 5,000 AVs was a reasonable fleet size to both operate an *on-demand SAMS with no shared rides*, and to answer the research problem in this study.

The 5,000 AVs travel at a fixed rate of 5 m/s (11 mph) because this is approximately the average taxi speed in Manhattan. The simulation assumes AVs take 15 seconds to drop off a user and 45 seconds to pick up a user.

Each simulation runs from 3:00am to 11:59pm (i.e. $T = 21$hours); however, user requests only enter the system between 3:00am and 10:30pm. The AVs finish picking up and dropping off users between 10:30pm and 11:59pm. This procedure ensures the fleet can serve all requests in all scenarios. In all scenarios, all the AVs are initially located (at 3:00am) in one location. This forces the AVs to reposition before the morning peak period.

The simulation time step $\Delta\tau$ is one second. The inter-decision interval I^d is 30 seconds, which is long enough to solve the optimization problem instances in this study. Allowing user requests to queue over 30 seconds also allows the fleet controller to make efficient AV-user assignments. The impact of I^d on the operational performance is another interesting research area; however, it is beyond the scope of this thesis.

The value of wait time c^{VOT} and the empty distance cost rate c^{ED} used in the objective function are \$21.6/hr. and \$0.3/km (\$0.48/mi), respectively. These values coincide with estimates of value of time in the literature and the U.S. governmental mileage rate (Internal Revenue Service, 2018). The reward for assigning an AV to open user request r^{asgn} is \$2.1/user. Given the empty distance cost rate c^{ED} of \$0.3/km, AVs will not be assigned to new user requests if the AVs are more than 7.0 km (minus $c^{VOT} \times w_i$) away from the new request. A smaller assignment reward value would increase user wait times but decrease empty fleet kilometers. The imbalance penalty c^{I} is \$1.5/user, indicating an empty AV will only be considered for repositioning to an imbalanced subregion, if the AV and subregion are less than 5.0 km away from each other. The parameter values chosen by an SAMS fleet operator will likely depend on how they want to position themselves in the market. If the SAMS fleet is concerned with user wait times they can choose larger values for value of wait time c^{VOT}, and the assignment reward r^{asgn}. If they are more concerned with offering low prices through keeping their operational costs down, they can decrease the assignment reward r^{asgn} and the imbalance penalty c^{I} to avoid empty miles.

In this study, the minimum imbalance parameter I^{min} is set to 1 vehicle. This parameter allows the fleet to control the aggressiveness of empty AV repositioning. Larger values of I^{min} should decrease empty repositioning miles while increasing average user wait times; however, this analysis is beyond the scope of this thesis. The prediction horizon h^{p} is 30 minutes.

Table 9-1: Parameter Settings that do not Vary across Scenarios

Parameter	Math Notation	Value	Units
Fleet Size		5000	vehicles
Vehicle Speed	v	5	meters/sec.
Drop-off Time		15	sec.
Pickup Time		45	sec.
Simulation Length	T	21	hours
Simulation Time Step	$\Delta\tau$	1	sec.
Inter-decision Interval	I^D	30	sec.
Value of Time	c^{VOT}	21.6	\$/hr.
Empty Distance Cost Rate	c^{EDCR}	0.3	\$/km
Assignment Reward	r^{asgn}	2.1	\$/user
Imbalance Penalty	c^I	1.5	\$/user
Minimum Imbalance	I^{min}	1	user
Prediction Horizon	h^p	30	minutes

Scenarios

Given the parameter values listed in the previous subsection, the computational analysis involves simulating the performance of a SAMS fleet under a variety of scenarios. The scenarios vary:

- Forecast type: Perfect and model forecasts
- Spatial partition for demand forecasts (side 1 length | side 2 length | area):
    - 2.83 km | 2.65 km | 7.49 km^2
    - 1.41 km | 1.32 km | 1.87 km^2
    - 0.71 km | 0.66 km | 0.47 km^2
    - 0.35 km | 0.33 km | 0.12 km^2
- Temporal aggregation for demand forecasts: 5-min, 30-min, and 60-min
- Request data: 30 days of taxi data from April 2016

In combination, this represents 2x4x3x30=720 simulations/scenarios. The analysis also includes 30 experiments for the no forecast/no AV repositioning case. In the scenarios without repositioning, the fleet solves the math program in Eqn. (46)-(49) every I^d.

9.7 Results

Demand Forecasting Results

This subsection presents statistical error measurements for demand forecasts across different spatio-temporal aggregation levels. The following two metrics are used to measure statistical error:

$$RMSRE = \frac{1}{N_T}\frac{1}{N_{nz}}\sum_{h=1}^{N_T}\sum_{k\in Z_{nz}}\sqrt{\left(\frac{X_k^h - Z_k^h}{Z_k^h}\right)^2} ,$$

$$sMAPE = \frac{1}{N_T}\frac{1}{N_{nz}}\sum_{h=1}^{N_T}\sum_{k\in Z_{nz}}\frac{|X_k^h - Z_k^h|}{X_k^h + Z_k^h + 1} ,$$

where N_T is the number of time intervals and Z_{nz} is the set of subregions containing demand ($N_{nz} = |Z_{nz}|$). X_k^h and Z_k^h are the forecasted number of requests in subregion k during time-interval h from the demand model and the perfect forecast, respectively.

Figure 9-5 displays the average statistical error values for all days in April 2016 for different aggregation levels. The figure shows that errors increase with both shorter time intervals and smaller subregions. This finding is consistent across the two error measures employed in this study – root mean squared relative error (RMSRE) and symmetric mean absolute percentage error (sMAPE). In general, the RMSE metric penalizes large individual errors between actual and observed demand ($X_k^h - Z_k^h$) more severely than the sMAPE metric. However, the trends in Figure 9-5 are similar for both error measures. The relationship between temporal aggregation as well as spatial aggregation and statistical forecast error appears to be non-linear. A rigorous analysis of these relationships requires more data points and is beyond the scope of this study.

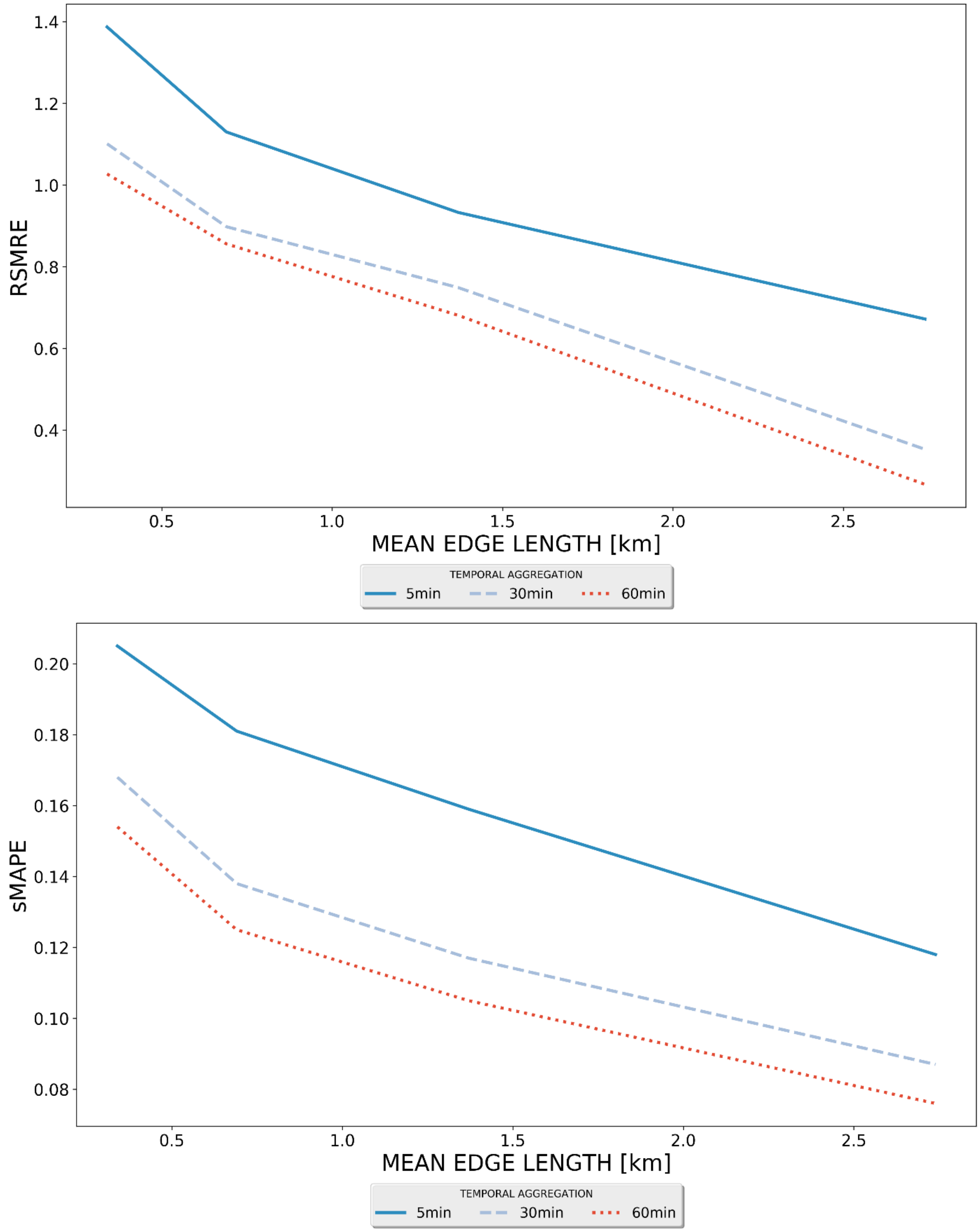


Figure 9-5: Average demand forecast errors for different spatial (x-axis) and temporal (line type) aggregation levels according to RSMRE (top) and sMAPE (bottom) error measures

SAMS Fleet Performance Results

This subsection presents the results of the computational analysis that was designed to evaluate and quantify the impact of spatio-temporal demand forecast aggregation on SAMS fleet performance. It includes two key performance metrics: average user wait time and the share of empty fleet miles.

Figure 9-6 displays average user wait time as a function of spatial aggregation and forecast type (the temporal aggregation level is 5 minutes in Figure 9-6). Each point on the model and perfect forecast lines represents the mean of thirty separate experiments (i.e. the 30 days of April 2016) for a single spatial aggregation level. As the scenarios without repositioning do not depend on forecast aggregation, the dotted line represents the mean of one set of 30 experiments.

There are several important findings displayed in Figure 9-6. First, average user wait time increases significantly with spatial aggregation; i.e. using more disaggregate demand forecasts and smaller subregions significantly improve fleet performance in terms of average user wait time. Second, it is not until demand forecasts are spatially highly disaggregate and subregions are small that a fleet using perfect demand forecasts begins to significantly outperform a fleet using model forecasts.

These first two findings provide some evidence to reject (and support) the hypothesis presented in this chapter. Although Figure 9-6 shows continued improvement at progressively more disaggregate levels, a comparison of the SAMS performance under perfect forecasts and model forecasts suggests that it becomes progressively more difficult to improve fleet performance at highly-disaggregate levels using model forecasts with significant errors. This indicates that

SAMS providers can benefit from improving demand forecast methods, especially at more spatially disaggregate levels.

Third, a fleet using no information about advanced requests and no empty AV repositioning outperforms a fleet using perfect demand forecasts in terms of average user wait times, when subregions are large in this study. This finding suggests that the SAMS operational strategy employed in this study is suboptimal in general, but especially when the service region is divided into large subregions. One potential method to improve the SAMS operational strategy includes making sure the AVs do not cluster in subregion centroids or at the edge of subregions when repositioning. The operational strategy could force the available/empty AVs to spread out within their current subregions. This would decrease the distance between new user requests and the available AVs in their subregion. This improvement would likely have the biggest positive impact when subregions are large. Additionally, adjusting the parameters in operational strategies objective function to emphasize reducing wait times and de-emphasize reducing empty fleet miles, may improve the average user wait times for large subregions.

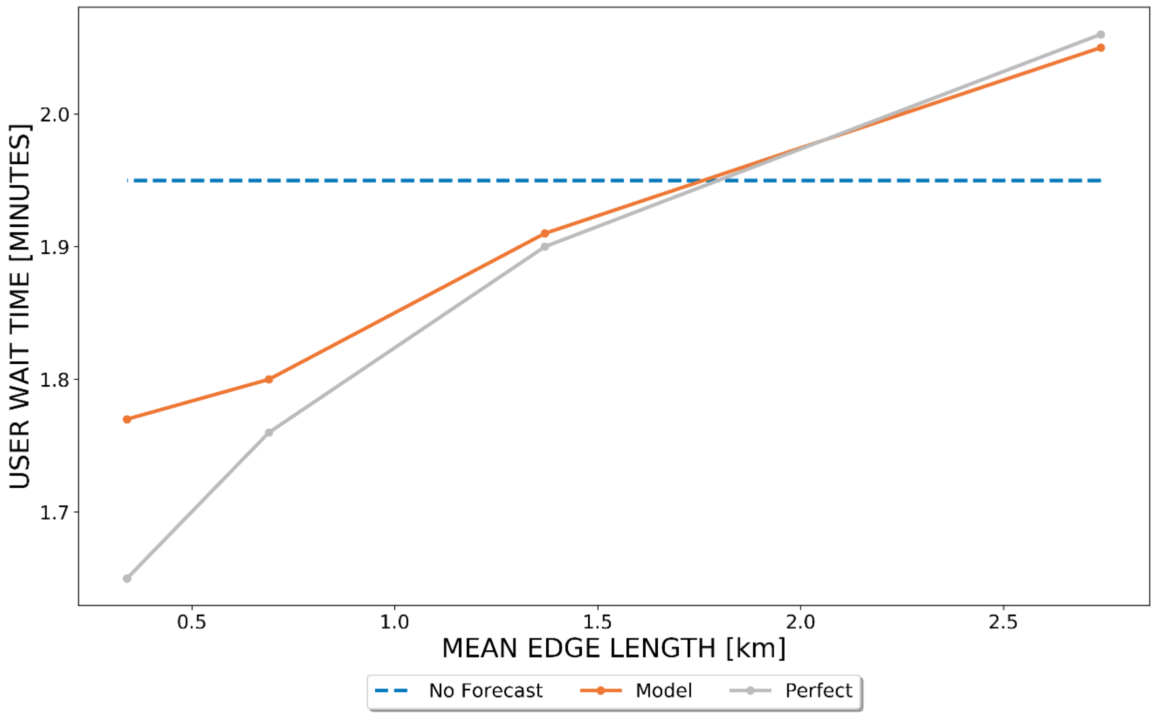


Figure 9-6 Average user wait time as a function of subregion edge length (x-axis) and forecast type (line color)

Figure 9-7 displays the percentage of fleet miles that are empty across different spatial aggregation levels. The solid lines at the top of the figure represent *total* empty miles; whereas, the small vertical dash lines in the middle represent empty *pickup* miles and the horizontal dashed lines at the bottom represent empty *repositioning* miles. *Total* empty miles are the summation of empty *pickup* miles and empty *repositioning* miles.

The results in Figure 9-7 are quite interesting, especially in the context of Figure 9-6. Once again, fleet performance (measured in total fleet miles) improves with more disaggregate demand forecasts and smaller subregions. This finding suggests that there is not a trade-off in terms of operational costs and service quality when choosing a spatial aggregation level; rather, more disaggregate forecasts and smaller subregions perform better across both metrics.

Additionally, Figure 9-7 indicates why/how more disaggregate forecasts and smaller subregions produce the shorter wait times in Figure 9-6. Empty *pickup* miles significantly decrease for smaller subregions, meaning AVs are positioned closer to new user requests when subregions are smaller, effectively decreasing user wait times. This significant decrease in empty *pickup* miles more than offsets the increase in empty *repositioning* miles for smaller subregions.

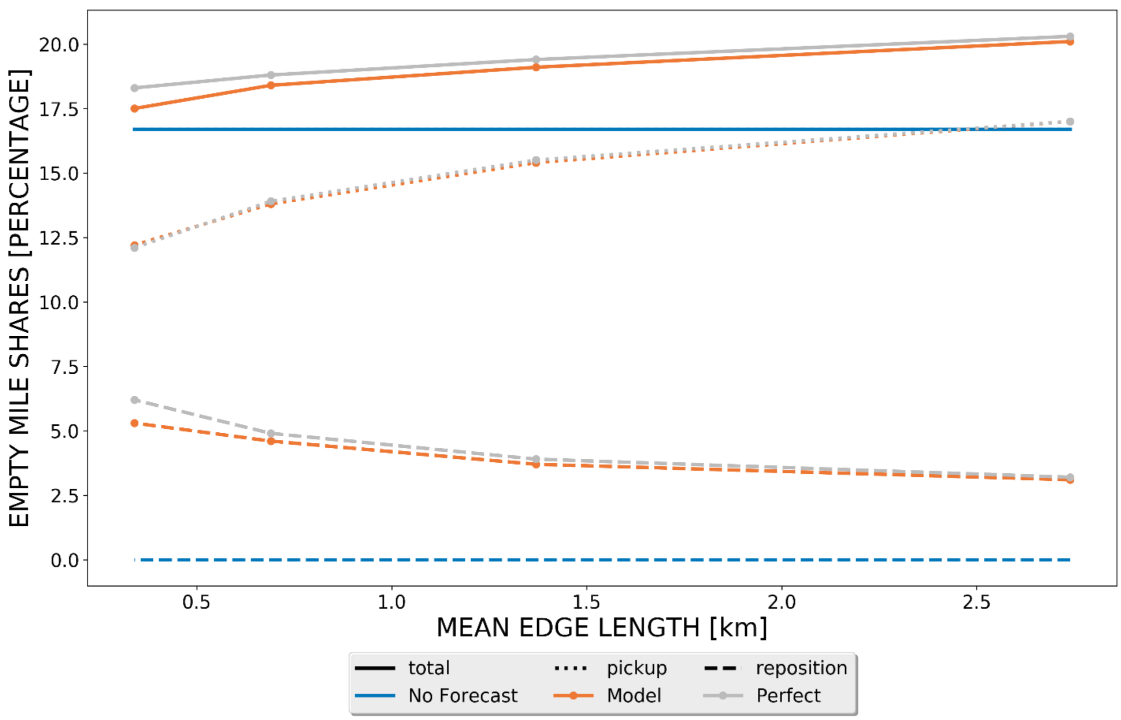


Figure 9-7 Empty fleet miles as a function of subregion edge length (x-axis), type of empty miles (line type), and forecast type (line color)

Table 9-2 shows the computational results in tabular form for all three temporal aggregation levels. The table indicates that for the SAMS operational strategy in this study (i) temporal aggregation level had minimal impact on fleet performance, and (ii) the relationship between spatial aggregation and fleet performance holds across temporal aggregation levels. This first result likely does not hold in general. In fact, it suggests that the SAMS operational strategy employed in this study fails to effectively use higher-resolution temporal forecasts to improve operational performance. Additional research is needed improve the SAMS operational strategies in this study to properly capture the temporal aspects of demand forecasts

Table 9-2: Complete SAMS Fleet Performance Results as a Function of Forecast Type, Spatial Aggregation, and Temporal Aggregation

Forecast Type	Edge Length (km)	Avg. User Wait Time (min)	Empty Pickup Miles Share	Empty Reposition Miles Share	Total Empty Miles Share
None	NA	1.95	16.7%	0.0%	16.7%
5-min Temporal Aggregation					
Model	2.7	2.05	17.0%	3.1%	20.1%
	1.35	1.91	15.4%	3.7%	19.1%
	0.68	1.80	13.8%	4.6%	18.4%
	0.34	1.77	12.2%	5.3%	17.5%
Perfect	2.7	2.06	17.0%	3.2%	20.3%
	1.35	1.90	15.5%	3.9%	19.4%
	0.68	1.76	13.9%	4.9%	18.8%
	0.34	1.65	12.1%	6.2%	18.3%
30-min Temporal Aggregation					
Model	2.7	2.05	17.0%	3.1%	20.1%
	1.35	1.91	15.4%	3.7%	19.1%
	0.68	1.80	13.8%	4.6%	18.4%
	0.34	1.78	12.2%	5.3%	17.5%
Perfect	2.7	2.06	17.0%	3.2%	20.2%
	1.35	1.91	15.5%	3.8%	19.3%
	0.68	1.79	13.9%	4.7%	18.6%
	0.34	1.74	12.2%	5.6%	17.7%
60-min Temporal Aggregation					
Model	2.7	2.05	17.0%	3.2%	20.1%
	1.35	1.91	15.4%	3.8%	19.1%
	0.68	1.80	13.8%	4.6%	18.4%
	0.34	1.77	12.2%	5.4%	17.5%
Perfect	2.7	2.06	17.0%	3.2%	20.3%
	1.35	1.91	15.5%	3.8%	19.4%
	0.68	1.78	14.0%	4.8%	18.7%
	0.34	1.69	12.2%	5.8%	18.0%

9.8 Conclusion

Summary and Implications

This study evaluates and quantifies the impact of spatio-temporal demand forecast aggregation on the operational performance of an *on-demand SAMS without shared rides*. This research problem combines two timely research areas, namely, forecasting demand for mobility services and developing strategies to dynamically operate SAMSs efficiently. The existing literature largely treats these problems independently despite their inherent interconnection. Hence, the research problem and methodological framework presented in this chapter have significant practical value to SAMS providers who need to forecast demand in order to efficiently operate their AV fleets.

The computational analysis illustrates that more disaggregate forecasts significantly improve SAMS fleet performance in terms of empty fleet miles *and* user wait times. As forecasts become more disaggregate, the SAMS fleet more effectively repositions AVs into smaller subregions, thereby decreasing average user wait times and empty *pickup* miles. The decrease in empty *pickup* miles more than offsets the increase in empty *repositioning* miles. Additionally, the results indicate that while demand forecast quality has little impact on fleet performance when spatial aggregation is high, as demand forecasts become more disaggregate, forecast quality begins to significantly impact operational performance.

These findings suggest that (i) there are significant benefits associated with dividing service areas into smaller subregions to forecast demand and reposition AVs, and (ii) improvements in demand forecasting methods, particularly for disaggregate spatial scales, can produce significant value to on-demand SAMSs in terms of operational performance.

Limitations and Future Work

This study presented a variety of challenges in terms of conducting a truly scientific analysis to test the study's hypothesis. The study design clearly defines the demand forecasting model, the SAMS operational strategy, the NYC taxi data, and the modeling assumptions. Nevertheless, the SAMS operational strategy (i.e. the assignment and repositioning algorithm) is not an optimal strategy because it is highly unlikely that an optimal strategy exists for such a highly-dynamic, stochastic, and large problem. Hence, there is no way to guarantee results will hold across SAMS operational strategies. Moreover, there is no way to guarantee the results will hold across different demand forecasting methods, and in different service areas.

This limitation suggests the research problem presented in this chapter along with the flexible methodological framework represent more significant scientific contributions than the results for one demand forecasting method, one set of taxi data, and one SAMS operational strategy. Future research should employ the methodological framework presented in this study, but use different SAMS operational strategies, demand forecasting methods, and different user request data to further test this study's hypothesis.

The 0.34-km edge length is the smallest spatial scale presented in this study due to computational constraints. Smaller edge lengths increase the computational time to solve the joint assignment-repositioning problem. Future work should improve computational performance and test more disaggregate spatial subregions. There is also room for improvement in terms of the SAMS operational strategy and its exploitation of the demand forecast output.

Another future research area of interest is the impact of demand forecast errors on SAMS operational performance. Testing different demand forecast models will result in a variation in demand forecast errors, that can be used to determine the relationship between demand forecast errors and SAMS operational performance. It is also possible to employ a single model, or perfect forecasts, and systematically create errors in the forecasts to answer this research question. In this study, forecast errors are also a function of the different demand data; i.e. the different days in the taxi data. A study design, which only varies forecast errors while keeping demand and aggregation level constant, could also highlight if any forecast error measure correlates better with fleet performance results.

Finally, future work can more effectively handle temporal components of the short-term demand forecasts within the repositioning strategy.

Chapter 10 Concluding Remarks

10.1 Summary and Contributions

This thesis focuses on addressing operational problems associated with on-demand shared-use AV mobility services (SAMSs), which are inherently dynamic and stochastic problems. The motivation for this topic arises from the rapid growth of ridesourcing companies (e.g. Uber and Lyft) and their impact on personal mobility and entire metropolitan transportation systems (Clewlow and Mishra, 2017; Rayle et al., 2016). Further motivation comes from the expected advent of fully-autonomous vehicles (AVs) and their eventual inclusion within shared-use mobility services. AVs should only increase the market share of shared-use mobility services via decreasing operational costs.

There are many research challenges surrounding SAMSs; however, the data analysis of taxi efficiency in Chapter 4 illustrates that without central operation, mobility services may operate inefficiently. Operating SAMS fleets more efficiently can improve service quality, reduce operational costs, and increase the competitiveness of mobility service providers. Moreover, operational efficiency affects fleet miles thereby impacting congestion, fuel consumption, and vehicle emissions. Hence, the operational problems, models, and analyses presented in this thesis can inform transportation policy makers who are interested in the impacts of SAMSs on individual mobility and metropolitan transportation systems.

The core contributions of this thesis include defining several SAMSs, delineating the underlying problems associated with operating these SAMSs, and presenting solution approaches to address the stochastic dynamic problems associated with operating on-demand SAMSs. The SAMSs presented in this thesis include the *on-demand SAMS without shared rides* (Chapter 6), the *on-demand shared-ride SAMS* (Chapter 7), and the *on-demand autonomous carsharing service* (Chapter 8). These on-demand SAMS operational problems represents new instances of stochastic dynamic vehicle routing problems. The combination of the SAMS operational problems' size, degree of dynamism, degree of urgency, spatial distribution of user requests, and short user pickup and drop-off times make the problem instances unique relative to the existing literature. Moreover, the solution approaches presented in this thesis for these unique stochastic dynamic vehicle routing problem instances utilize the mathematical programming formulation of the assignment (i.e. bipartite matching) problem formulation. Because the linear relaxation of the assignment problem returns integer solutions, the formulation is quite useful in the context of highly-dynamic and stochastic problems where large problems need to be solved repeatedly.

Additional contributions presented in this thesis include the taxonomy of VRPs developed to classify SAMS operational problem classes and problem instances (Chapter 2); the operational efficiency analysis of the Chicago taxi fleet (Chapter 4); the in-depth sensitivity analysis on the impact of the maximum in-vehicle user detour time parameter on customer service quality and operational efficiency for the *on-demand shared-ride SAMS* (Chapter 7); and the methodology developed to quantify the impact of spatio-temporal demand forecast aggregation on the performance of an *on-demand SAMS without shared rides* (Chapter 9);

10.2 Applications

The research presented in this thesis has several important application areas. The first, and most apparent, application area is the operation of SAMSs in the future. Hopefully, fleet operators gain insight from the conceptualization of SAMS operational problem instances in this thesis, as well as find the models and solution approaches useful for different applications. The taxonomy of VRPs in Chapter 2 provides an overview of important considerations in the design of mobility services for service providers. The mathematical models and solution approaches in this thesis are also meant to be implementable for service providers. The integration of short-term demand forecasting and SAMS operational models in Chapter 9 should be particularly useful for SAMS providers.

The second application is the modeling of urban transportation systems. Transportation planners employ transportation system models in order to understand the impacts of different planning and policy changes on metropolitan transportation systems. For example: how will congestion pricing impact traffic congestion in urban areas? Or, how will AVs and SAMS impact travel behavior, demand for transit, demand for roadways, and roadway capacity? However, most existing transportation system models do not incorporate mobility services. Hence, the models and solution methods presented in this thesis provide a strong basis for modeling SAMS fleets in transportation networks.

10.3 Future Research Areas

Modeling and optimizing SAMS fleets that serve passengers is a relatively new area of research that requires further study. One of the biggest challenges involves incorporating the advanced SAMS fleet operational strategies within a traffic simulation model. In reality, SAMS fleets will significantly impact traffic congestion and traffic congestion will significantly impact the operation of SAMS fleets. Most of the modeling frameworks in the literature, including the one presented in this thesis, ignore the interrelationships between SAMS fleets and traffic congestion. Properly capturing this interrelationship is very important for transportation planning and policy questions, as well as SAMS operational problems. More broadly, the notion of vehicle route choices and user-equilibrium on road networks needs to be reconsidered in era of shared mobility.

The taxonomy in Chapter 2 illustrates that there are numerous SAMS operational problems that still need to be addressed. Although Alonso-Mora et al. (2017) present important research on the shared-ride problem, there is still room for advancements. Another challenging SAMS operational problem of interest involves modeling a SAMS with advanced *and* immediate traveler demand requests. The advanced demand requests will have tight and strict time-windows; whereas, the immediate demand requests will still want to be assigned to an AV immediately and picked up within a few minutes.

Other open research areas related to SAMSs include incorporating pricing into the dynamic fleet modeling framework (Chen and Kockelman, 2016; Figliozzi et al., 2007; Sayarshad and Chow, 2015) and allowing travelers to accept or reject price and wait time offers from SAMS fleet operators. Incorporating pricing and allowing users to reject SAMS offers within an SAMS modeling framework requires the development and integration of behavioral models.

Finally, there needs to be a real focus on developing robust solution algorithms for the operational problems associated with SAMSs. Although trucking, taxi, and other vehicle fleets currently use real-time control algorithms to provide *decision support* to vehicle dispatchers and/or drivers, with AVs, control algorithms will need to *make decisions* rather than support decision makers. The difference is important and needs to be reflected in the solution algorithms and modeling frameworks.

REFERENCES

Agatz, N., Erera, A., Savelsbergh, M., Wang, X., 2012. Optimization for dynamic ride-sharing: A review. European Journal of Operational Research 223, 295–303. doi:10.1016/j.ejor.2012.05.028

Alonso-Mora, J., Samaranayake, S., Wallar, A., Frazzoli, E., Rus, D., 2017. On-demand high-capacity ride-sharing via dynamic trip-vehicle assignment. Proceedings of the National Academy of Sciences of the United States of America 114, 462–467. doi:10.1073/pnas.1611675114

Andersson, T., Värbrand, P., 2007. Decision support tools for ambulance dispatch and relocation. Journal of the Operational Research Society 58, 195–201. doi:10.1057/palgrave.jors.2602174

Archetti, C., Speranza, M.G., Vigo, D., 2014. Chapter 10: Vehicle Routing Problems with Profits, in: Vehicle Routing. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 273–297. doi:10.1137/1.9781611973594.ch10

Bektas, T., Repoussis, P.P., Tarantilis, C.D., 2014. Chapter 11: Dynamic Vehicle Routing Problems, in: Vehicle Routing. Society for Industrial and Applied Mathematics, Philadelphia, PA, pp. 299–347. doi:10.1137/1.9781611973594.ch11

Berbeglia, G., Cordeau, J.-F., Laporte, G., 2010. Dynamic pickup and delivery problems. European Journal of Operational Research 202, 8–15. doi:10.1016/j.ejor.2009.04.024

Bertsimas, D., de Boer, S., 2005. Simulation-Based Booking Limits for Airline Revenue Management. Operations Research 53, 90–106. doi:10.1287/opre.1040.0164

Bilegan, I.C., Brotcorne, L., Feillet, D., Hayel, Y., 2015. Revenue management for rail container transportation. EURO Journal on Transportation and Logistics 4, 261–283. doi:10.1007/s13676-014-0051-7

Bischoff, J., Maciejewski, M., 2016. Simulation of City-wide Replacement of Private Cars with Autonomous Taxis in Berlin, in: Procedia Computer Science. doi:10.1016/j.procs.2016.04.121

Bodin, L., Golden, B., 1981. Classification in vehicle routing and scheduling. Networks 11, 97–108. doi:10.1002/net.3230110204

Boesch, P.M., Ciari, F., Axhausen, K.W., 2016. Autonomous Vehicle Fleet Sizes Required to Serve Different Levels of Demand. Transportation Research Record: Journal of the Transportation Research Board 2542, 111–119. doi:10.3141/2542-13

Brotcorne, L., Laporte, G., Semet, F., 2003. Ambulance location and relocation models.

European Journal of Operational Research 147, 451–463. doi:10.1016/S0377-2217(02)00364-8

Brown, T., 2016. Matchmaking in Lyft Line - Part 3 [WWW Document]. Lyft Engineering.

Burns, L., Jordan, W., Scarborough, B., 2013. Transforming Personal Mobility. New York.

Chaniotakis, E., Antoniou, C., Pereira, F., 2016. Mapping Social media for transportation studies. IEEE Intelligent Systems 31, 64–70. doi:10.1109/MIS.2016.98

Chen, T.D., Kockelman, K.M., 2016. Management of a Shared Autonomous Electric Vehicle Fleet: Implications of Pricing Schemes. Transportation Research Record: Journal of the Transportation Research Board 2572, 37–46. doi:10.3141/2572-05

Chen, T.D., Kockelman, K.M., Hanna, J.P., 2016. Operations of a shared, autonomous, electric vehicle fleet: Implications of vehicle & charging infrastructure decisions. Transportation Research Part A: Policy and Practice 94, 243–254. doi:10.1016/j.tra.2016.08.020

Chen, Y., Hyland, M., Wilbur, M.P., Mahmassani, H.S., 2018. Characterization of Taxi Fleet Operational Networks and Vehicle Efficiency: A Chicago Case Study. Transportation Research Record.

Chicago Data Portal, 2017. Taxi Trips. Chicago.

Clewlow, R., Mishra, G.S., 2017. Disruptive Transportation: The Adoption, Utilization, and Impacts of Ride-Hailing in the United States.

Dandl, F., Bogenberger, K., 2018a. Booking Processes in Autonomous Carsharing and Taxi Systems, in: Transportation Research Arena. Vienna, Austria.

Dandl, F., Bogenberger, K., 2018b. Comparing Future Autonomous Electric Taxis with an Existing Free-Floating Carsharing System. accepted to IEEE ITS Transactions 1–11.

Dandl, F., Bracher, B., Bogenberger, K., 2017. Microsimulation of an autonomous taxi-system in Munich, in: 2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS). IEEE, pp. 833–838. doi:10.1109/MTITS.2017.8005628

Dantzig, G.B., Ramser, J.H., 1959. The Truck Dispatching Problem. Management Science 6, 80–91. doi:10.1287/mnsc.6.1.80

Downs, A., 2004. Why Traffic Congestion Is Here to Stay. . . and Will Get Worse. ACCESS 19–25. doi:https://www.accessmagazine.org/wp-content/uploads/sites/7/2016/07/Access-25-04-Traffic-Congestion-is-Here-to-Stay.pdf

Eksioglu, B., Vural, A.V., Reisman, A., 2009. The vehicle routing problem: A taxonomic review. Computers & Industrial Engineering 57, 1472–1483.

doi:10.1016/j.cie.2009.05.009

Fagnant, D.J., Kockelman, K.M., 2016. Dynamic ride-sharing and fleet sizing for a system of shared autonomous vehicles in Austin, Texas. Transportation 1–16. doi:10.1007/s11116-016-9729-z

Fagnant, D.J., Kockelman, K.M., 2014. The travel and environmental implications of shared autonomous vehicles, using agent-based model scenarios. Transportation Research Part C: Emerging Technologies 40, 1–13. doi:10.1016/j.trc.2013.12.001

Fagnant, D.J., Kockelman, K.M., Bansal, P., 2015. Operations of Shared Autonomous Vehicle Fleet for Austin, Texas, Market. Transportation Research Record: Journal of the Transportation Research Board 2536, 98–106. doi:10.3141/2536-12

Figliozzi, M.A., Mahmassani, H.S., Jaillet, P., 2007. Pricing in Dynamic Vehicle Routing Problems. Transportation Science 41, 302–318.

Fleischmann, B., Gnutzmann, S., Sandvoß, E., 2004. Dynamic Vehicle Routing Based on Online Traffic Information. Transportation Science 38, 420–433.

Ford, H., 2012. Shared Autonomous Taxis: Implementing an Efficient Alternative to Automobile Dependency. Princeton University.

Frantzeskakis, L.F., Powell, W.B., 1990. A Successive Linear Approximation Procedure for Stochastic, Dynamic Vehicle Allocation Problems. Transportation Science 24, 40–57. doi:10.1287/trsc.24.1.40

Furuhata, M., Dessouky, M., Ordóñez, F., Brunet, M.-E., Wang, X., Koenig, S., 2013. Ridesharing: The state-of-the-art and future directions. Transportation Research Part B: Methodological 57, 28–46. doi:10.1016/j.trb.2013.08.012

Gendreau, M., Laporte, G., Séguin, R., 1996. Stochastic vehicle routing. European Journal of Operational Research 88, 3–12. doi:10.1016/0377-2217(95)00050-X

Gendreau, M., Laporte, G., Semet, F., 2001. A dynamic model and parallel tabu search heuristic for real-time ambulance relocation. Parallel Computing 27, 1641–1653. doi:10.1016/S0167-8191(01)00103-X

Goodson, J.C., Ohlmann, J.W., Thomas, B.W., 2013. Rollout Policies for Dynamic Solutions to the Multivehicle Routing Problem with Stochastic Demand and Duration Limits. Operations Research 61, 138–154. doi:10.1287/opre.1120.1127

Google, 2017. Google Maps API [WWW Document]. URL https://developers.google.com/maps/documentation/distance-matrix/start.

Haggag, K., McManus, B., Paci, G., 2017. Learning by Driving: Productivity Improvements by New York City Taxi Drivers. American Economic Journal: Applied Economics 9, 70–95.

doi:10.1257/app.20150059

Haghani, A., Yang, S., 2007. Real-Time Emergency Response Fleet Deployment: Concepts, Systems, Simulation & Case Studies, in: Dynamic Fleet Management. Springer US, Boston, MA, pp. 133–162. doi:10.1007/978-0-387-71722-7_7

Hanna, J.P., Albert, M., Chen, D., Stone, P., 2016. Minimum Cost Matching for Autonomous Carsharing, in: Proceedings of the 9th IFAC Symposium on Intelligent Autonomous Vehicles. Leipzig, Germany.

Hörl, S., Ruch, C., Becker, F., Frazzoli, E., Axhausen, K.W., 2017. Fleet control algorithms for automated mobility : A simulation assessment for Zurich, in: TRB 2018 Annual Meeting. doi:10.3929/ETHZ-B-000175260

Hosni, H., Naoum-Sawaya, J., Artail, H., 2014. The shared-taxi problem: Formulation and solution methods. Transportation Research Part B: Methodological 70, 303–318. doi:10.1016/j.trb.2014.09.011

Hyland, M., Mahmassani, H.S., 2018. Dynamic autonomous vehicle fleet operations: Optimization-based strategies to assign AVs to immediate traveler demand requests. Transportation Research Part C: Emerging Technologies 92, 278–297. doi:10.1016/j.trc.2018.05.003

Hyland, M.F., Mahmassani, H.S., 2017. Taxonomy of Shared Autonomous Vehicle Fleet Management Problems to Inform Future Transportation Mobility. Transportation Research Record: Journal of the Transportation Research Board 2653, 26–34. doi:10.3141/2653-04

IBM, 2017. Descriptive, predictive, prescriptive: Transforming asset and facilities management with analytics. Somers.

Ichoua, S., Gendreau, M., Potvin, J.-Y., 2006. Exploiting Knowledge About Future Demands for Real-Time Vehicle Dispatching. Transportation Science 40, 211–225. doi:10.1287/trsc.1050.0114

Iglesias, R., Rossi, F., Wang, K., Hallac, D., Leskovec, J., Pavone, M., 2017. Data-Driven Model Predictive Control of Autonomous Mobility-on-Demand Systems.

Ihler, A., Hutchins, J., Smyth, P., 2006. Adaptive event detection with time-varying poisson processes. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD ’06 207. doi:10.1145/1150402.1150428

Internal Revenue Service, 2018. Standard Mileage Rates [WWW Document]. URL https://www.irs.gov/tax-professionals/standard-mileage-rates (accessed 10.16.18).

International Transport Forum, 2015. Urban Mobility System Upgrade: How shared self-driving cars could change city traffic.

Jaillet, P., Wagner, M.R., 2006. Online Routing Problems: Value of Advanced Information as Improved Competitive Ratios. Transportation Science 40, 200–210.

Kiam Tian Seow, Nam Hai Dang, Der-Horng Lee, 2010. A Collaborative Multiagent Taxi-Dispatch System. IEEE Transactions on Automation Science and Engineering 7, 607–616. doi:10.1109/TASE.2009.2028577

King, D.A., Saldarriaga, J.F., 2017. Spatial Regulation of Taxicab Services: Measuring Excess Travel from New York City, in: Transportation Research Board 96th Annual Meeting. Transportation Research Board, Washington DC, United States.

Krueger, R., Rashidi, T.H., Rose, J.M., 2016. Preferences for shared autonomous vehicles. Transportation Research Part C: Emerging Technologies 69, 343–355. doi:10.1016/j.trc.2016.06.015

Lahyani, R., Khemakhem, M., Semet, F., 2015. Rich vehicle routing problems: From a taxonomy to a definition. European Journal of Operational Research 241, 1–14. doi:10.1016/J.EJOR.2014.07.048

Laporte, G., 1992. The vehicle routing problem: An overview of exact and approximate algorithms. European Journal of Operational Research 59, 345–358. doi:10.1016/0377-2217(92)90192-C

Larsen, A., Madsen, O.B.G., Solomon, M.M., 2002. Partially Dynamic Vehicle Routing-Models and Algorithms. Journal of the Operational Research Society 53, 637–646.

Lee, A., Savelsbergh, M., 2015. Dynamic ridesharing: Is there a role for dedicated drivers? Transportation Research Part B: Methodological 81. doi:10.1016/j.trb.2015.02.013

Lee, D.-H., Wang, H., Cheu, R., Teo, S., 2004. Taxi Dispatch System Based on Current Demands and Real-Time Traffic Conditions. Transportation Research Record 1882, 193–200.

Lee, S., 2012. The role of centrality in ambulance dispatching. Decision Support Systems 54, 282–291. doi:10.1016/j.dss.2012.05.036

Levin, M.W., Fritz, H., Boyles, S.D., 2017a. On Optimizing Reservation-Based Intersection Controls. IEEE Transactions on Intelligent Transportation Systems 18, 505–515. doi:10.1109/TITS.2016.2574948

Levin, M.W., Kockelman, K.M., Boyles, S.D., Li, T., 2017b. A general framework for modeling shared autonomous vehicles with dynamic network-loading and dynamic ride-sharing application. Computers, Environment and Urban Systems 64, 373–383. doi:10.1016/j.compenvurbsys.2017.04.006

Liu, X., Gong, L., Gong, Y., Liu, Y., 2015. Revealing travel patterns and city structure with taxi trip data. Journal of Transport Geography 43, 78–90. doi:10.1016/j.jtrangeo.2015.01.016

Lu, Q., Dessouky, M.M., 2006. A new insertion-based construction heuristic for solving the pickup and delivery problem with time windows. European Journal of Operational Research 175, 672–687. doi:10.1016/j.ejor.2005.05.012

Lund, K., Madsen, O.B.G., Rygaard, J.M., 1996. Vehicle routing with varying degree of dynamism.

Ma, S., Zheng, Y., Wolfson, O., 2015. Real-Time City-Scale Taxi Ridesharing. IEEE Transactions on Knowledge and Data Engineering 27, 1782–1795. doi:10.1109/TKDE.2014.2334313

Maciejewski, M., Bischoff, J., Nagel, K., 2016. An Assignment-Based Approach to Efficient Real-Time City-Scale Taxi Dispatching. IEEE Intelligent Systems 31, 68–77. doi:10.1109/MIS.2016.2

Maciejewski, M., Nagel, K., 2013. Simulation and dynamic optimization of taxi services in MATSim.

Makridakis, S., 1988. Metaforecasting : Ways of improving forecasting accuracy and usefulness. International Journal of Forecasting 4, 467–491.

Martinez, L.M., Viegas, J.M., 2017. Assessing the impacts of deploying a shared self-driving urban mobility system: An agent-based model applied to the city of Lisbon, Portugal. International Journal of Transportation Science and Technology 6, 13–27. doi:10.1016/j.ijtst.2017.05.005

McGill, J.I., van Ryzin, G.J., 1999. Revenue Management: Research Overview and Prospects. Transportation Science 33, 233–256. doi:10.1287/trsc.33.2.233

Mckenzie, B., Rapino, M., 2011. Commuting in the United States: 2009.

Moreira-Matias, L., Gama, J., Ferreira, M., Mendes-Moreira, J., Damas, L., 2013. Predicting taxi-passenger demand using streaming data. IEEE Transactions on Intelligent Transportation Systems 14, 1393–1402. doi:10.1109/TITS.2013.2262376

Müller, J., Bogenberger, K., 2015. Time series analysis of booking data of a free-floating carsharing system in Berlin. Transportation Research Procedia 10, 345–354. doi:10.1016/j.trpro.2015.09.084

Nasrollahzadeh, A.A., Khademi, A., Mayorga, M.E., 2018. Real-Time Ambulance Dispatching and Relocation. Manufacturing & Service Operations Management 20, 467–480. doi:10.1287/msom.2017.0649

NYC Taxi & Limousine Commission, 2017. TLC Trip Record Data [WWW Document]. URL http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml (accessed 7.24.18).

Pavone, M., Smith, S.L., Frazzoli, E., Rus, D., 2012. Robotic load balancing for mobility-on-demand systems. International Journal of Robotics Research 31, 839–854. doi:10.1177/0278364912444766

Pavone, M., Smith, S.L., Frazzoli, E., Rus, D., 2011. Load Balancing for Mobility-on-Demand Systems, in: Robotics: Science and Systems 2011. Los Angeles.

Pillac, V., Gendreau, M., Guéret, C., Medaglia, A.L., 2013. A review of dynamic vehicle routing problems. European Journal of Operational Research 225, 1–11. doi:10.1016/j.ejor.2012.08.015

Powell, W.B., 2011. Approximate Dynamic Programming, Second. ed. Hoboken.

Powell, W.B., 1996. A Stochastic Formulation of the Dynamic Assignment Problem, with an Application to Truckload Motor Carriers. Transportation Science 30, 195–219. doi:10.1287/trsc.30.3.195

Psaraftis, H.N., 1983. An Exact Algorithm for the Single Vehicle Many-to-Many Dial-A-Ride Problem with Time Windows. http://dx.doi.org/10.1287/trsc.17.3.351 17, 351–357.

Psaraftis, H.N., 1980. A Dynamic Programming Solution to the Single Vehicle Many-to-Many Immediate Request Dial-a-Ride Problem. Transportation Science 14, 130–154. doi:10.1287/trsc.14.2.130

Psaraftis, H.N., Wen, M., Kontovas, C.A., 2016. Dynamic vehicle routing problems: Three decades and counting. Networks 67, 3–31. doi:10.1002/net.21628

Pureza, V., Laporte, G., 2008. Waiting and Buffering Strategies for the Dynamic Pickup and Delivery Problem with Time Windows. INFOR: Information Systems and Operational Research 46, 165–176. doi:10.3138/infor.46.3.165

Puterman, M.L., 2014. Markov decision processes : discrete stochastic dynamic programming. Wiley-Interscience.

Qian, X., Ukkusuri, S. V., 2015. Spatial variation of the urban taxi ridership using GPS data. Applied Geography 59, 31–42. doi:10.1016/j.apgeog.2015.02.011

Rayle, L., Dai, D., Chan, N., Cervero, R., Shaheen, S., 2016. Just a better taxi? A survey-based comparison of taxis, transit, and ridesourcing services in San Francisco. Transport Policy 45, 168–178. doi:10.1016/j.tranpol.2015.10.004

Regan, A., Mahmassani, H., Jaillet, P., 1996. Dynamic Decision Making for Commercial Fleet Operations Using Real-Time Information. Transportation Research Record: Journal of the Transportation Research Board 1537, 91–97. doi:10.3141/1537-13

Regan, A.C., Mahmassani, H.S., Jaillet, P., 1995. Improving Efficiency of Commercial Vehicle Operations Using Real-time Information; Potential Uses and Assignment Strategies. Transportation Research Record 188–198.

Rigole, P.-J., 2014. Study of a Shared Autonomous Vehicles Based Mobility Solution in Stockholm.

Ritzinger, U., Puchinger, J., Hartl, R.F., 2015. A survey on dynamic and stochastic vehicle routing problems. International Journal of Production Research 54, 215–231. doi:10.1080/00207543.2015.1043403

Ropke, S., Cordeau, J.-F., 2009. Branch and Cut and Price for the Pickup and Delivery Problem with Time Windows. Transportation Science 43, 267–286. doi:10.1287/trsc.1090.0272

Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S.H., Ratti, C., 2014. Quantifying the benefits of vehicle pooling with shareability networks. Proceedings of the National Academy of Sciences of the United States of America 111, 13290–4. doi:10.1073/pnas.1403657111

Santos, D.O., Xavier, E.C., 2015. Taxi and Ride Sharing: A Dynamic Dial-a-Ride Problem with Money as an Incentive. Expert Systems with Applications 42, 6728–6737. doi:10.1016/j.eswa.2015.04.060

Savelsbergh, M.W.P., Sol, M., 1995. The General Pickup and Delivery Problem. Transportation Science 29, 17–29. doi:10.1287/trsc.29.1.17

Sayarshad, H.R., Chow, J.Y.J., 2017. Non-myopic relocation of idle mobility-on-demand vehicles as a dynamic location-allocation-queueing problem. Transportation Research Part E: Logistics and Transportation Review 106, 60–77. doi:10.1016/j.tre.2017.08.003

Sayarshad, H.R., Chow, J.Y.J., 2016. Survey and empirical evaluation of nonhomogeneous arrival process models with taxi data. Journal of Advanced Transportation 50, 1275–1294. doi:10.1002/atr.1401

Sayarshad, H.R., Chow, J.Y.J., 2015. A scalable non-myopic dynamic dial-a-ride and pricing problem. Transportation Research Part B: Methodological 81, 539–554. doi:10.1016/j.trb.2015.06.008

Schmid, V., 2012. Solving the dynamic ambulance relocation and dispatching problem using approximate dynamic programming. European Journal of Operational Research 219, 611–621. doi:10.1016/j.ejor.2011.10.043

Schmid, V., Doerner, K.F., 2010. Ambulance location and relocation problems with time-dependent travel times. European Journal of Operational Research 207, 1293–1303. doi:10.1016/j.ejor.2010.06.033

Schneider, T.W., 2016. Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance [WWW Document]. URL http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance/ (accessed 7.27.17).

Sheridan, P.K., Gluck, E., Guan, Q., Pickles, T., Balcıog˘lu, B., Benhabib, B., 2013. The dynamic nearest neighbor policy for the multi-vehicle pick-up and delivery problem. Transportation Research Part A: Policy and Practice 49, 178–194. doi:10.1016/j.tra.2013.01.032

Spieser, K., Samaranayake, S., Gruel, W., Frazzoli, E., 2016. Shared-vehicle Mobility-on-Demand Systems: A Fleet Operator's Guide to Rebalancing Empty Vehicles, in: 95th Annual Meeting of the Transportation Research Board. Washington D.C.

Spieser, K., Treleaven, K., Zhang, R., Frazzoli, E., Pavone, M., 2014. Toward a Systematic Approach to the Design and Evaluation of Automated Mobility-on-Demand Systems: A Case Study in Singapore, in: Meyer, G., Beiker, S. (Eds.), Road Vehicle Automation, Lecture Notes in Mobility. Springer International Publishing, pp. 229–245. doi:10.1007/978-3-319-05990-7

Talebpour, A., Mahmassani, H.S., 2016. Influence of connected and autonomous vehicles on traffic flow stability and throughput. Transportation Research Part C: Emerging Technologies 71, 143–163. doi:10.1016/J.TRC.2016.07.007

Thomas, B.W., 2007. Waiting Strategies for Anticipating Service Requests from Known Customer Locations. Transportation Science 41, 319–331. doi:10.1287/trsc.1060.0183

Tjokroamidjojo, D., Kutanoglu, E., Taylor, G.D., 2006. Quantifying the value of advance load information in truckload trucking. Transportation Research Part E: Logistics and Transportation Review 42, 340–357. doi:10.1016/j.tre.2005.01.001

Tong, Y., Chen, Y., Zhou, Z., Chen, L., Wang, J., Yang, Q., Ye, J., Lv, W., 2017. The Simpler The Better. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17 1653–1662. doi:10.1145/3097983.3098018

Toth, P., Vigo, D. (Eds.), 2002. The Vehicle Routing Problem. Society for Industrial and Applied Mathematics. doi:10.1137/1.9780898718515

Tsao, M., Iglesias, R., Pavone, M., 2018. Stochastic Model Predictive Control for Autonomous Mobility on Demand.

Ulmer, M., Goodson, J., Mattfeld, D., Thomas, B., 2017. Route-based Markov decision processes for dynamic vehicle routing problems.

Ulmer, M., Thomas, B.W., Campbell, A.M., Ulmer, M.W., Woyak, N., 2017. The Restaurant Meal Delivery Problem: Dynamic Pick-Up and Delivery with Deadlines and Random

Ready Times.

Ulmer, M.W., 2017. Approximate Dynamic Programming for Dynamic Vehicle Routing, 61st ed, Operations Research/Computer Science Interfaces Series. Springer.

Ulmer, M.W., 2016. Anticipation vs. Reoptimization for Dynamic Vehicle Routing with Stochastic Requests.

van Lon, R.R.S., Ferrante, E., Turgut, A.E., Wenseleers, T., Berghe, G. Vanden, Holvoet, T., 2016. Measures of dynamism and urgency in logistics. European Journal of Operational Research 253, 614–624. doi:10.1016/J.EJOR.2016.03.021

Viegas, J., Martinez, L., 2016. Shared Mobility: Innovation for Liveable Cities.

Voccia, S.A., Melissa Campbell, A., Thomas, B.W., 2017. The Same-Day Delivery Problem for Online Purchases. Transportation Science trsc.2016.0732. doi:10.1287/trsc.2016.0732

Wang, R., 2015. The stops made by commuters: evidence from the 2009 US National Household Travel Survey. Journal of Transport Geography 47, 109–118. doi:10.1016/J.JTRANGEO.2014.11.005

Waymo, 2017. Waymo's fully self-driving vehicles are here [WWW Document]. Medium. URL https://medium.com/waymo/with-waymo-in-the-drivers-seat-fully-self-driving-vehicles-can-transform-the-way-we-get-around-75e9622e829a (accessed 11.7.17).

Weikl, S., Bogenberger, K., 2015. A practice-ready relocation model for free-floating carsharing systems with electric vehicles - Mesoscopic approach and field trial results. Transportation Research Part C: Emerging Technologies 57, 206–223. doi:10.1016/j.trc.2015.06.024

Wilson, N.H.M., Colvin, N.J., 1977. Computer control of the Rochester dial-a-ride system, Technical Report R77-31. Cambridge, Massachusetts.

Wingfield, N., 2017. Automakers Race to Get Ahead of Silicon Valley on Car-Sharing. New York Times.

Winter, K., Cats, O., van Arem, B., Martens, K., 2017. Impact of relocation strategies for a fleet of shared automated vehicles on service efficiency, effectiveness and externalities. IEEE, pp. 844–849. doi:10.1109/MTITS.2017.8005630

Yang, C., Gonzales, E.J., 2017. Modeling Taxi Demand and Supply in New York City Using Large-Scale Taxi GPS Data, in: Thakuriah, P., Tilahun, N., Zellner, M. (Eds.), Seeing Cities Through Big Data: Research, Methods and Applications in Urban Informatics. Springer, pp. 405–425. doi:10.1007/978-3-319-40902-3_22

Yang, J., Jaillet, P., Mahmassani, H., 2004. Real-Time Multivehicle Truckload Pickup and Delivery Problems. Transportation Science 38, 135–148.

Yang, J., Jaillet, P., Mahmassani, H., 1999. On-Line Algorithms for Truck Fleet Assignment and Scheduling Under Real-Time Information. Transportation Research Record: Journal of the Transportation Research Board 1667, 107–113. doi:10.3141/1667-13

Zachariah, J., Gao, J., Kornhauser, A., Mufti, T., 2014. Uncongested Mobility for All: A Proposal for an Area Wide Autonomous Taxi System in New Jersey, in: Transportation Research Board 93rd Annual Meeting.

Zha, L., Yin, Y., Xu, Z., 2018. Geometric matching and spatial pricing in ride-sourcing markets. Transportation Research Part C: Emerging Technologies 92, 58–75. doi:10.1016/J.TRC.2018.04.015

Zhan, X., Qian, X., Ukkusuri, S. V., 2016. A Graph-Based Approach to Measuring the Efficiency of an Urban Taxi Service System. IEEE Transactions on Intelligent Transportation Systems 17, 2479–2489. doi:10.1109/TITS.2016.2521862

Zhan, X., Qian, X., Ukkusuri, S. V, 2014. Measuring the Efficiency of Urban Taxi Service System, in: Proc. The 3rd International Workshop on Urban Computing. New York, NY, USA.

Zhang, R., Pavone, M., 2016. Control of robotic mobility-on-demand systems: A queueing-theoretical perspective. International Journal of Robotics Research 35, 186–203. doi:10.1177/0278364915581863

Zhang, R., Rossi, F., Pavone, M., 2018. Analysis, Control, and Evaluation of Mobility-on-Demand Systems: a Queueing Theoretical Approach. IEEE Transactions on Control of Network Systems 1–1. doi:10.1109/TCNS.2018.2800403

Zhang, R., Rossi, F., Pavone, M., 2015. Model Predictive Control of Autonomous Mobility-on-Demand Systems. doi:10.1109/ICRA.2016.7487272

Zhang, W., Guhathakurta, S., Fang, J., Zhang, G., 2015a. The Performance and Benefits of a Shared Autonomous Vehicles Based Dynamic Ridesharing System: An Agent-Based Simulation Approach, in: Transportation Research Board 94th Annual Meeting.

Zhang, W., Guhathakurta, S., Fang, J., Zhang, G., 2015b. Exploring the Impact of Shared Autonomous Vehicles on Urban Parking Demand: An Agent-based Simulation Approach. Sustainable Cities and Society 19, 34–45. doi:10.1016/j.scs.2015.07.006

Zhong, C., Batty, M., Manley, E., Wang, J., Wang, Z., Chen, F., Schmitt, G., 2016. Variability in regularity: Mining temporal mobility patterns in London, Singapore and Beijing using smart-card data. PLoS ONE 11. doi:10.1371/journal.pone.0149222

Zhu, S., Kornhauser, A., 2017. The Interplay Between Fleet Size, Level-of-Service and Empty Vehicle Repositioning Strategies in Large-Scale, Shared-Ride Autonomous Taxi Mobility-on-Demand Scenarios, in: Transportation Research Board 96th Annual

Meeting. Transportation Research Board, Washington DC.

Zotteri, G., Kalchschmidt, M., Caniato, F., 2005. The impact of aggregation level on forecasting performance. International Journal of Production Economics 93–94, 479–491. doi:10.1016/J.IJPE.2004.06.044